Robust Prediction and Biomarker Discovery in Rare Cancers Using Interpretable Machine Learning

Dhurka Rohini Madasamy

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2025

Committee:

Wooyoung Kim

Munehiro Fukuda

Bill Erdly

Program Authorized to Offer Degree:

Computer Science & Software Engineering

University of Washington

Abstract

Robust Prediction and Biomarker Discovery in Rare Cancers Using Interpretable Machine Learning

Dhurka Rohini Madasamy

Chair of the Supervisory Committee:
Wooyoung Kim
Department of Computing and Software Systems

Rare cancers such as Glioblastoma Multiforme (GBM, a rare brain cancer) pose persistent challenges in computational oncology due to limited data, biological noise, and difficulty in isolating disease-specific molecular signatures. Based on these constraints, this work began with the expectation that rare-cancer models would perform poorly. However, machine learning approaches on genomic data achieved unexpectedly strong accuracy, motivating investigation into whether this separability reflected genuine biology or artifactual signal. This thesis develops an interpretable machine learning framework that evaluates predictive robustness and isolates biologically meaningful biomarkers under extreme imbalance. Cascade Learning systematically removes broad cancer pathways and reveals biomarkers uniquely associated with the rare cancer, while SHAP-based interpretability aligns these genes with experimentally reported glioma biology. Complementary Tab2Image visualizations provide spatial confirmation of class separability, strengthening biological trust in the learned signal. Overall, this work provides a robust, biologically grounded, and ethically aligned pathway for rare-cancer biomarker discovery that emphasizes transparency, fairness, and accountability in scarce-data environments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

PREDICTIVE MODELING: A technique used to build models that can predict future outcomes based on patterns learned from historical data.

MACHINE LEARNING: A field of artificial intelligence focused on developing algorithms that learn from data and make predictions or decisions without being explicitly programd.

DATA PREPROCESSING: The process of cleaning and transforming raw data into a format suitable for analysis, including handling missing values, normalizing features, and encoding categorical variables.

FEATURE ENGINEERING: The creation, selection, or transformation of variables (features) to improve machine learning model performance.

CLASSIFICATION: A modeling task in which an algorithm predicts a categorical label. In this thesis, the task involves predicting whether a sample belongs to rare GBM or a comparison class.

LOGISTIC REGRESSION (LR): A linear and interpretable classification model used to evaluate whether rare cancer signals are linearly separable from other classes.

RANDOM FOREST: An ensemble learning algorithm built from multiple decision trees. It is robust to noise and provides useful measures of feature importance.

XGBOOST: A high-performance gradient boosting algorithm that builds an ensemble of trees using iterative, loss-minimizing updates. It often achieves state-of-the-art accuracy on tabular data.

ROC CURVE: A plot showing the trade-off between true positive rate and false positive rate across varying classification thresholds.

AREA UNDER THE CURVE (AUC): A scalar summary of the ROC curve that measures how well a classifier separates positive and negative classes.

INTERPRETABILITY: The degree to which a human can understand the reasoning behind a machine learning model's predictions.

FEATURE IMPORTANCE: A measure of how strongly a given feature (e.g., a gene) contributes to a model's predictions.

BIOMARKER: A molecular feature, such as a gene or transcript, that consistently distinguishes one biological condition from another.

CASCADE LEARNING: A two-stage modeling strategy that removes general cancer signals before training a second model to isolate rare-cancer-specific biomarkers.

CODING GENES: Genes that encode proteins. These features were used as a biologically interpretable subset for classification and Tab2Img transformations.

CLASS IMBALANCE: A condition in which one class (such as rare GBM samples) appears far less frequently than comparison classes, requiring specialized sampling strategies.

DEEPINSIGHT: A Tab2Img transformation that converts tabular gene expression data into spatially meaningful images compatible with convolutional neural networks.

F1 SCORE: A metric combining precision and recall, particularly useful when class imbalance is severe, as in rare cancer detection.

GBM (GLIOBLASTOMA MULTIFORME): A rare brain cancer that serves as the positive class across all scopes analyzed in this thesis.

GTEX (GENOTYPE-TISSUE EXPRESSION PROJECT): A large database of RNA-seq profiles from normal tissues. It provides the healthy brain samples used in this study.

IMBALANCED-LEARN (IMBLEARN): A Python toolkit offering resampling strategies such as SMOTE and RUS to address class imbalance in high-dimensional datasets.

LGG (LOWER GRADE GLIOMA): A common brain cancer used as a comparison group when evaluating rare GBM classification.

MCC (MATTHEWS CORRELATION COEFFICIENT): A balanced metric that evaluates classification quality even under extreme class imbalance.

MINMAXSCALER: A normalization method that rescales features to the 0–1 range, essential for Tab2Img transformations and fair model comparison.

PCA (PRINCIPAL COMPONENT ANALYSIS): An unsupervised dimensionality reduction technique that visualizes separability between rare GBM, common cancers, and normal tissue based on gene expression.

RARE CANCER: A cancer type with limited patient representation in genomic databases. In this thesis, the rare cancer of interest is GBM.

RUS (RANDOM UNDERSAMPLING): A resampling method that reduces the number of majority-class samples to mitigate class imbalance.

SHAP (SHAPLEY ADDITIVE EXPLANATIONS): A model-agnostic interpretability method based on cooperative game theory that quantifies each gene's contribution to a model's predictions.

SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE): A method that synthesizes new minority-class samples using feature-space interpolation, helping balance rare-cancer datasets.

SUPERTML: A Tab2Img algorithm that places features into fixed spatial templates based on ranked importance to create structured image representations.

TAB2IMG (TAB-TO-IMAGE): A family of algorithms that convert high-dimensional tabular gene expression vectors into images to enable visual and CNN-based interpretation.

TCGA (THE CANCER GENOME ATLAS): A large cancer genomics repository providing expression data for common cancers and rare GBM.

GENE EXPRESSION MATRIX: A structured table where rows represent samples and columns represent gene expression measurements, forming the input for all modeling tasks.

PHENOTYPE METADATA: Sample-level annotations (e.g., tissue type, study origin, gender) used for filtering, leakage detection, and scope construction.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my project advisor, Dr. Wooyoung Kim, for her unwavering support, guidance, and invaluable insights throughout the entire process of researching and writing this project. I am also indebted to my project committee members, Dr. Munehiro Fukuda and Dr. Bill Erdly, for their constructive feedback and thoughtful suggestions that significantly enriched the quality of this research. Their expertise and time devoted to reviewing and discussing my work are genuinely appreciated.

I would like to extend special thanks to William Selke, whose development of the Tab2Image framework formed an essential component of this project, and to Jeremy Newton, with whom I collaborated closely in the early stages of this work. We exchanged ideas, aligned experimental approaches, and supported each other throughout our parallel thesis efforts, and I am deeply grateful for his collegiality and insight.

I extend my heartfelt thanks to my husband and kids for their continuous support, understanding, and encouragement. Their belief in my abilities and their unwavering support sustained me during the challenges of graduate study.

I am grateful to my family, friends, and peers who provided camaraderie, academic discussion, and a supportive community throughout this journey. Their friendship made the academic experience both enjoyable and enriching.

Finally, I want to express my appreciation to all the individuals who, in various ways, contributed to the completion of this master's project. Your support has been invaluable, and I am sincerely thankful for your contributions.

Thank you all for being an integral part of this academic endeavor.

DEDICATION

To my dear husband Kali, my sons Nimal and Muhil, and my beloved brother and sisters.

Chapter 1

INTRODUCTION

This chapter establishes the scientific and ethical context for rare cancer classification, outlines the motivation for developing an interpretable modeling framework, and introduces the cascade-based approach used to refine biologically grounded biomarkers.

1.1 Context and Motivation

Rare diseases are defined by their low prevalence, affecting only small segments of the population. The Orphanet database identifies a rare disorder as one affecting fewer than one in two thousand individuals in Europe [1], and in the United States a disease is classified as rare if it affects fewer than two hundred thousand people. Although individually uncommon, these disorders collectively impact an estimated three hundred to four hundred million people worldwide [2].

These low numbers directly influence machine-learning applications. Limited representation reduces statistical power, fragmented cohorts elevate privacy risk, and incomplete datasets challenge fairness [3]. As several studies note[4, 5], rare cancer datasets are particularly challenging: many samples are not publicly released because they could be reidentified, and institutions often hesitate to share data. As a result, research tends to focus on common cancers, leaving rare cancers both scientifically underrepresented and ethically disadvantaged.

Machine learning has become a powerful tool in cancer genomics because it can uncover patterns in high-dimensional gene expression data[6]. However, these advances have benefited common cancers disproportionately, where large repositories such as TCGA provide abundant, well-annotated samples supporting reproducible model development [7]. Rare cancers remain underrepresented in genomic research due to the intrinsic difficulty of acquiring

clinically verified expression data [3]. Their scarcity is not simply demographic: structural, political, and privacy constraints limit data visibility and sharing. Reidentification risks restrict public release of small cohorts, institutional barriers slow data exchange, and incentive structures favor diseases with larger treatment markets. These forces create fragmented datasets that weaken statistical signal and increase the likelihood of biased outcomes [8, 9].

International frameworks from WHO, UNESCO, and the European Commission emphasize that trustworthy biomedical AI must support fairness, accountability, transparency, and reliability [10]. Rare cancers highlight the consequences when these principles are unmet. In such settings, performative accuracy masks representational inequity and low-resource biology is overshadowed by better-funded disease groups.

Glioblastoma (GBM), the topic of this thesis, represents this tension clearly. Despite being a highly lethal cancer, GBM remains underrepresented in major datasets, and its molecular diversity complicates analysis. Addressing these gaps motivated the development of an interpretable pipeline. Methods such as SHAP, PCA, and Tab2Image were incorporated not simply for visualization but to ensure that model decisions could be explained and evaluated for biological consistency.

Glioblastoma (GBM), the rare cancer analyzed in this thesis, illustrates these tensions. GBM profiles are often missing from public repositories due to heightened privacy risk, and cohort sizes are typically small even within TCGA and related initiatives [11]. The motivation for this work is therefore grounded not only in technical relevance but in addressing representational inequity. Interpretability methods such as SHAP, PCA visualization, and Tab2Image analysis are incorporated to amplify rare cancer signal, make model reasoning auditable, and strengthen epistemic accountability [12, 13].

Cascade Learning further reflects this stance. Prior studies demonstrate that multistage frameworks can refine classifier specificity by filtering confounding biological signals before second-stage learning [14]. Applying this strategy here prevents dominant expression patterns in more common cancers from overshadowing the GBM signature. The methodology becomes both computational and ethical, ensuring that underrepresented biology remains

visible.

Early experiments reinforced this motivation. Conventional transfer learning approaches that reused pretrained cancer models were evaluated, aligning with reports that transfer learning frequently collapses in biologically disjoint domains [15, 16]. These models yielded trivial performance for GBM, whereas interpretable tabular models unexpectedly achieved near-perfect accuracy. This divergence strengthened the decision to pursue transparent, domain-aware modeling rather than pretrained embeddings. The contrast is illustrated in Figure B.1.

1.2 Problem Statement

Rare cancers experience structural inequities because limited funding, privacy barriers, and restricted data sharing constrain research capacity. These forces weaken predictive generalization and increase the likelihood of biased or clinically unsafe behavior.

This cycle raises questions of accountability. If a model built from inequitable data produces unfair or unstable decisions, responsibility lies not only with engineers but with systems that devalued rare cancer evidence.

The methodology in this thesis is shaped to counter these dynamics. Resampling techniques reduce imbalance, interpretability exposes model reasoning, multi scope evaluation avoids dependence on a single split, and Cascade Learning filters broad cancer signals so GBM specific features remain recoverable. The central challenge is twofold. We must achieve predictive stability despite scarcity and isolate biologically meaningful features rather than broad tumor effects.

Given the severe class imbalance, biological noise, and limited availability of GBM samples, the initial expectation was that machine-learning models would perform weakly or unstably. However, early experiments produced near-perfect accuracy, raising concerns about whether the separability was genuine or driven by hidden artifacts. This led to a refined hypothesis: if the predictive signal reflects true GBM biology, then the key genes should recur across models, remain detectable even after filtering broad cancer pathways through

Cascade Learning, and show consistency with biomarkers reported in glioma research.

1.3 Research Question, Motivation, and Approach

The research question emerges from this tension between expectation and observation. Does GBM possess a sufficiently distinct transcriptional signature for robust classification, and if so, how can we verify that its detectability reflects biological specificity rather than leakage or confounding?

Initial expectations assumed weak performance due to imbalance and heterogeneity. Instead, interpretable models performed unexpectedly well, necessitating investigation into why GBM was so separable and how to validate the credibility of its signal.

This thesis integrates three strategies: robustness evaluation across multiple biological scopes, interpretability driven biomarker discovery using SHAP and cross model recurrence, and Cascade Learning to filter dominant tumor signals so only rare cancer biology remains. The objective is not merely classification but explanation. We aim to understand the origin of separability by comparing our computational results with known glioma biomarkers.

1.4 Thesis Contributions

This work makes three contributions to rare cancer machine learning.

First, it demonstrates that GBM exhibits a stable and separable molecular signature across diverse modeling contexts. This reveals that GBM signal is deeply embedded in transcriptomic structure.

Second, it introduces an ethically aligned Cascade Learning framework that reduces interference from common cancer biology and sharpens biomarker specificity.

Third, it generates a reproducible consensus biomarker signature validated through cross model recurrence, SHAP explanations, and alignment with published glioma markers. The predictive strength of the models reflects authentic biology rather than artifact.

Together, these contributions form a transparent and ethically aligned pathway for rare cancer biomarker discovery integrating stability, interpretability, and biological validation.

Chapter 2

RELATED WORK

This chapter reviews the literature on rare cancer prediction, interpretability, biomarker discovery, and staged learning. These foundations contextualize the methodological decisions made in this thesis and highlight the gaps that motivate its design.

2.1 Machine Learning in Rare Cancer Prediction

Machine learning has become integral to cancer genomics, where high-dimensional gene-expression data facilitate tumor classification and biomarker inference [6]. Classical models such as Logistic Regression, Random Forest, and XGBoost remain effective because they are robust to noise, interpretable, and well suited to structured transcriptomic data [8]. Previous work has demonstrated reliable preprocessing practices and shown that stable signatures can emerge when cohorts are sufficiently large [13].

However, most reported findings are derived from cancers well represented in repositories such as TCGA, while rare cancers remain comparatively neglected [17]. Their small sample sizes, severe imbalance, and high molecular diversity reduce statistical power and increase susceptibility to modeling artifacts. Studies note unstable predictions on minority classes and degraded performance in unbalanced settings [18]. Sampling techniques such as SMOTE and undersampling offer partial mitigation but risk distorting biological signal. Comparative analyses therefore recommend evaluating models in multiple sampling settings rather than assuming that a single approach is universally applicable [19].However, the stability of predictions across varying biological scopes is rarely evaluated in prior work, and this thesis fills that gap.

Deep learning introduces an additional limitation. Although neural models perform well

on large datasets, multiple evaluations show degraded behavior under scarce or imbalanced conditions. Zhang et al. report unstable gradients, overfitting, and poor generalization for neural architectures in low-sample settings [16]. Gene-expression benchmarks indicate that deep learning surpasses classical models only when abundant training data exist; when cohort sizes shrink, performance advantages disappear or deteriorate [15]. These observations reinforce the continued relevance of interpretable, classical models in low-resource biomedical contexts.

2.2 Interpretable Machine Learning for Transparency

Interpretability methods such as SHAP, permutation importance, and feature rankings help reveal how models make decisions and support credible biomarker discovery [12]. Nonetheless, interpretability is often presented as a secondary add-on rather than a requirement for safety-critical biomedical systems, despite ethical AI guidance emphasising fairness, accountability, and transparency [20]. Rare cancers heighten these concerns because limited representation increases the risk of misleading patterns and unseen bias. Integrating interpretability throughout the modeling process therefore becomes necessary rather than optional, which motivates its central role in this thesis.

2.2.1 Biomarker Discovery and Feature Instability

Biomarker discovery pipelines commonly rely on methods such as LASSO, univariate filtering, or model-based importance scoring. However, multiple studies report that gene selection can be unstable when datasets are high dimensional, scarce, or imbalanced [21, 22]. Small changes in sampling, splitting, or preprocessing can produce divergent gene rankings, weakening confidence in biological generalisability [23]. Rare cancers intensify this risk because shared tumor signals often overshadow phenotype-specific effects, further eroding reproducibility [3]. Although several reviews recommend validating biomarkers across models, resampling regimes, and cohorts [21], very few studies systematically implement such assessments. The consensus-based approach adopted in this thesis responds directly

to this limitation by tracking recurrence across modeling contexts to strengthen biological credibility.

2.3 Multi-Stage and Cascade Learning Approaches

Multi-stage frameworks have been explored to improve classifier specificity by filtering confounding biological signals[14]. Cascade learning trains an initial model to capture broad cancer pathways, removes associated features, and trains a second-stage model on the refined feature set. Despite this promise, cascade-based filtering remains rarely applied to rare cancers, where benefits may be greatest. Moreover, conventional transfer learning pipelines trained on common cancers do not reliably transfer to rare cancers. Hanczar et al. report that pretrained embeddings yield inconsistent outcomes in minority tumor types and may collapse to trivial behavior when domain signals are mismatched [15]. These observations strengthen the rationale for domain-aware staged learning methods that progressively refine feature spaces.

2.4 Summary

Across the literature, rare cancers remain underrepresented in machine learning studies, deep learning models exhibit instability under scarce conditions [16], and transfer learning approaches often fail to adapt to minority tumor contexts [15]. Robustness across class scopes and sampling strategies is seldom evaluated, and biomarker stability is rarely validated despite its importance. Interpretability is frequently detached from ethical considerations, even though transparent reasoning is crucial for responsible biomedical deployment [20].

The framework developed in this thesis addresses these gaps through a robustness-focused pipeline, the integration of interpretability as both methodological and ethical necessity, and the application of Cascade Learning to isolate signatures specific to rare cancers. Through these advances, the thesis strengthens the link between computational rigor and responsible machine learning practice in rare cancer genomics.

Chapter 3

METHODOLOGY

This chapter presents the complete methodology used to evaluate the predictive robustness of the Glioblastoma Multiforme (GBM) gene-expression signature and to isolate rare cancer–specific biomarkers. It introduces the analytical pipeline, data processing procedures, experimental design, model architectures, the Cascade Learning framework, and interpretability analyses including SHAP and Tab-to-Image integration.

3.1 Overview

The methodology follows a staged design intended to evaluate whether the GBM rare cancer signal remains stable across progressively challenging settings. It combines dataset construction across four biological comparison scopes, interpretable modeling under multiple sampling regimes, and refinement through Cascade Learning and interpretability analysis. The following sections describe each component of this pipeline.

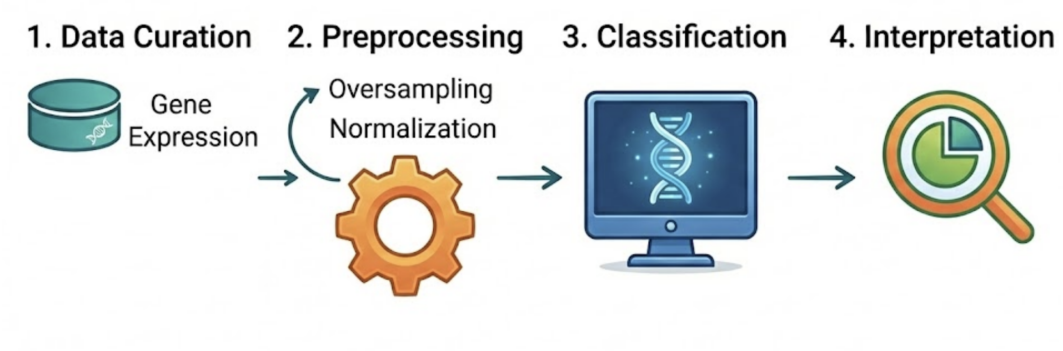

Figure 3.1: Overview of the complete machine-learning workflow.

3.2 Architecture and Technologies

All experiments were implemented in Python using a Jupyter Notebook environment, which enabled iterative statistical analysis, visual inspection of model behavior, and reproducible workflow execution. Pandas and NumPy supported dataset construction and numerical operations, while scikit-learn provided a unified framework for preprocessing, model development, and evaluation. Logistic Regression and Random Forest were implemented through scikit-learn, and XGBoost was introduced as a complementary gradient-boosting architecture capable of capturing non-linear relationships in transcriptomic data.

Class imbalance was handled using the Imbalanced-learn extension, where SMOTE was applied to generate synthetic minority-class observations and mitigate extreme rarity of GBM samples. Model interpretability relied on the SHAP framework, which quantified the contribution of each gene to prediction decisions and enabled biomarker validation. visualizations, including PCA maps, confusion matrices, and feature-attribution plots, were produced using Matplotlib and Seaborn to ensure consistent representation of analytical results. Together, these components formed a reproducible, scalable, and transparent computational ecosystem for evaluating predictive behavior and isolating biologically meaningful rare-cancer signal.

3.3 Gene Expression Data Sources and Provenance

This study uses harmonized bulk RNA–sequencing matrices curated through publicly accessible repositories widely adopted in computational oncology. Glioblastoma Multiforme (GBM) samples were obtained from The Cancer Genome Atlas (TCGA), which provides centrally sequenced and quality-controlled tumor transcriptomes under standardized bioinformatic processing pipelines. Normal brain expression profiles were sourced from the Genotype-Tissue Expression (GTEx) consortium, a population-scale reference dataset regarded as the benchmark for healthy tissue transcriptional baselines because TCGA does not include normal brain samples. Additional cancer comparators, including non-brain malignancies, were drawn from integration portals hosted through the UCSC Xena Browser, which aligns TCGA,

GTEx, and TARGET datasets under uniform normalization, schema validation, and access controls.

Gene expression captures transcript abundance across tens of thousands of genes and encodes pathway activity, cellular heterogeneity, and tumor phenotype variation, making it suitable for biomarker discovery and classification. Its high dimensionality and molecular noise, however, make rare-cancer settings particularly fragile because the absence of abundant cohorts heightens statistical variance and privacy sensitivity. Public repositories enforce anonymization and sample-sharing constraints precisely to reduce re-identification risk, which in turn limits dataset size and shapes the investigative space available for rare-disease machine learning.

To ensure reproducibility, all matrices were downloaded from UCSC Xena, screened for missingness, aligned to common identifiers, scaled to mitigate batch effects, and filtered to remove low-variance genes before modeling. The resulting dataset consists of GBM as the rare cancer of interest, Lower-Grade Glioma (LGG) as the biologically proximate common comparator, GTEx-derived normal brain tissue, and non-brain cancers used only in the broadest pan-cancer scope. While these repositories provide "clinical" metadata files, they contain only dataset descriptors such as tissue origin, sample type, primary site, and study source rather than true patient characteristics. Because these variables act as label synonyms rather than meaningful predictors, no clinical-feature modeling was conducted and all analyses rely exclusively on gene-expression inputs.

3.4 Dataset Size Summary

After integrating TCGA, GTEx, and TARGET repositories, the unified expression matrix contained 60,498 gene-expression features and six metadata descriptors. In total, 25,185 samples were available for analysis, of which 7,242 represented brain tissue or brain tumors derived exclusively from TCGA and GTEx. This brain-specific subset includes normal brain samples, Lower Grade Glioma (LGG) as a common cancer comparator, and Glioblastoma Multiforme (GBM) as the rare cancer of interest. Table 3.1 summarises the cohort compo-

sition.

Table 3.1: Dataset composition and feature dimensions

Dataset Component	**Samples**	**Features**
GTEx Expression Matrix	19,131	60,498
Merged TCGA–GTEx–TARGET Matrix	25,185	60,498
Brain Subset	7,242	60,498
Brain Normal (GTEx)	6,548	60,498
Brain Common Cancer – LGG (TCGA)	523	60,498
Brain Rare Cancer – GBM (TCGA)	171	60,498
All Other Cancers (TCGA + TARGET)	25,014	60,498

GBM represents only 171 cases among 7,242 brain samples, or roughly 2.36 percent of the brain cohort, illustrating its statistical rarity. This imbalance motivated the use of sampling techniques such as SMOTE and Random Undersampling, as well as the adoption of a multi-scope evaluation design to ensure that observed separability was not an artifact of skewed class frequencies.

Beyond the brain subset, the dataset includes a large and biologically diverse set of non-brain cancers predominantly sourced from TCGA. These samples become the negative class in the Rare-versus-All scope, reflecting the practical reality that rare-cancer biomarkers must remain identifiable even when contrasted against unrelated malignancies. This distribution mirrors the challenges of rare-cancer research more broadly, where rare cases are deeply dwarfed by heterogeneous biological backgrounds, making robust preprocessing and interpretability essential for reliable conclusions.

3.5 Exploratory Data Analysis (EDA)

Exploratory analysis highlighted both the structural imbalance and biological diversity present in the merged TCGA, GTEx, and TARGET dataset. As shown in Figure 3.2, GBM constitutes only a small fraction of all brain-related samples, reinforcing the need for sampling strategies such as SMOTE and Random Undersampling to stabilise learning under extreme rarity.

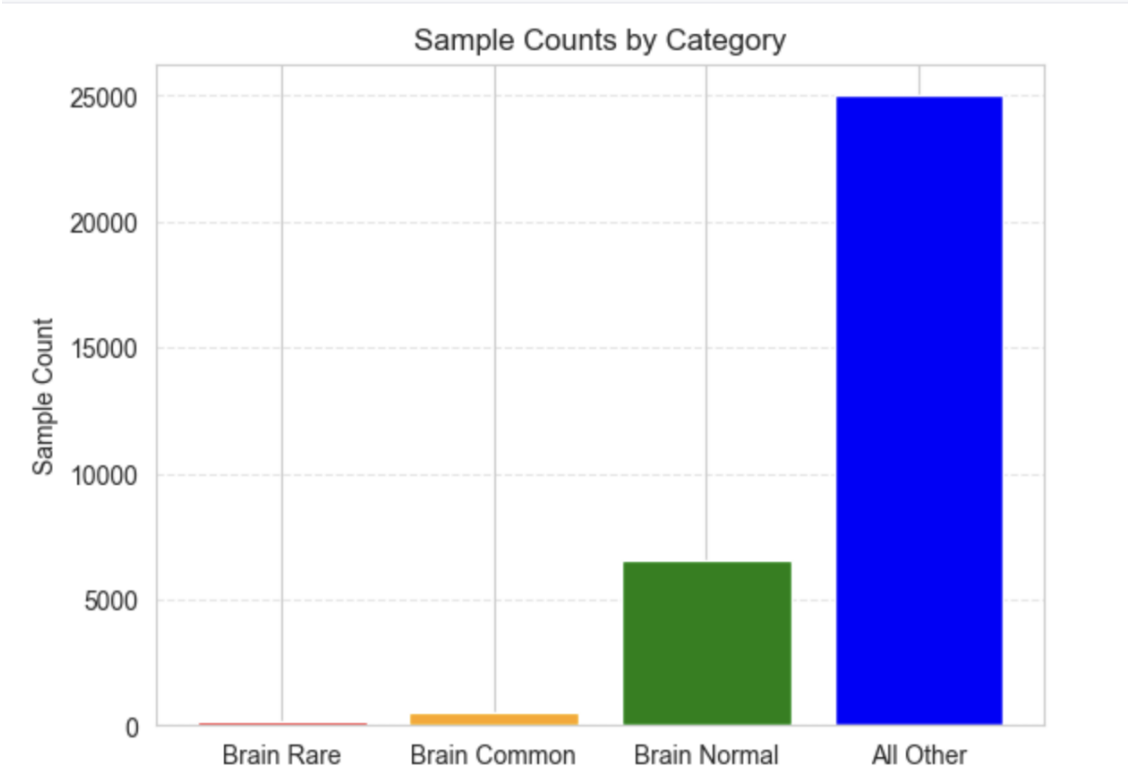

Figure 3.2: Distribution of rare, common, and normal samples showing significant imbalance.

Additional summarization confirmed that imbalance extends beyond tumor subtype. Figure 3.3 illustrates the contrast between tumor-rich TCGA cohorts and normal GTEx profiles in the Rare versus Normal setting. Across the full dataset, substantial variation in tissue and disease origin was observed during pre-analysis inspection, indicating that rare-cancer classi-

fication must be performed within a heterogeneous biological landscape. TARGET samples appear in these global distributions but are excluded from brain-specific experiments; they are included only in the pan-cancer comparison in Scope 4.

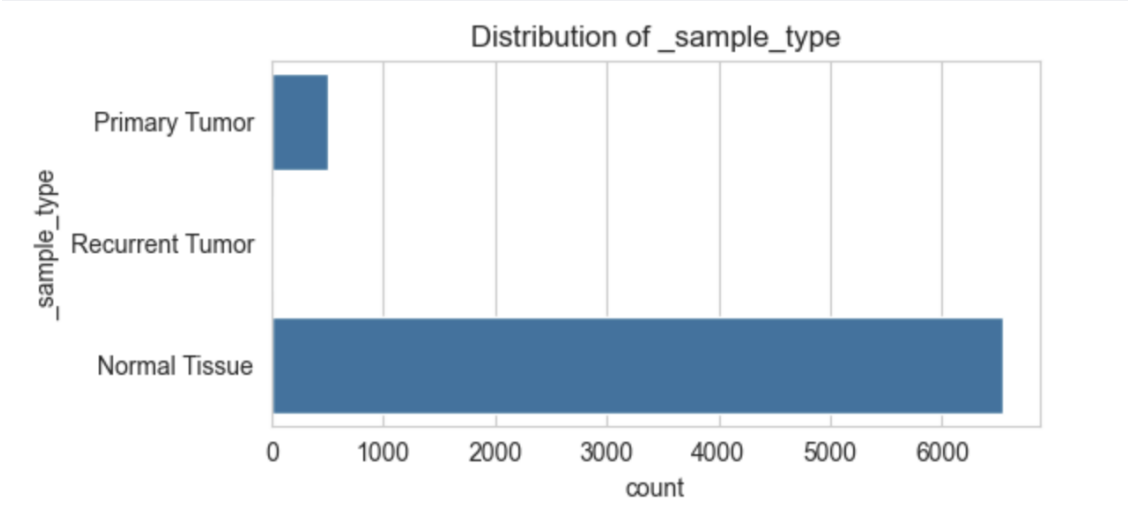

Figure 3.3: Distribution of sample types (tumor versus normal) in Scope 1.

Taken together, these exploratory findings show that rare-cancer signatures must be identified against a backdrop of severe imbalance and transcriptomic diversity. This complexity motivates the multi-scope experimental design and the balancing, filtering, and interpretability strategies adopted throughout the thesis.

3.6 Data Preprocessing

A standardized preprocessing workflow ensured that modeling relied solely on biologically meaningful inputs while preventing any form of data leakage.

Metadata fields that directly reflected class identity, such as *detailed_category* and *primary disease or tissue*, were removed because they were synonymous with the labels. The *_study* attribute was excluded for a similar reason: it identifies dataset origin (TCGA, GTEx, or TARGET) and would otherwise create an artificial shortcut between normal and tumor

samples.

Missing entries in remaining descriptors, including _gender, were imputed using scikit-learn's `SimpleImputer` under a most-frequent strategy. All imputation and scaling occurred inside the training pipeline to avoid contamination between training and validation data.

To enable biological comparison, two gene-expression feature sets were retained: the full transcriptome of approximately 60,000 genes and a protein-coding subset of roughly 20,000 genes. Parallel modeling across these spaces allowed assessment of whether rare-cancer separability depended on genome-wide information or remained evident when restricted to coding transcripts.

3.7 Experimental Design for Robustness Evaluation

To assess whether GBM exhibits a stable molecular signature, performance was evaluated across four classification scopes that progressively increase in biological difficulty. The scopes, summarized in Table 3.2, range from comparisons against normal brain tissue to the most heterogeneous contrast against all other cancers.

Table 3.2: Definition of the four classification scopes used in this study.

Scope	**Description**	**Positive Class**	**Negative Class**
1	Rare vs Normal	Rare cancer (GBM)	Normal brain tissue
2	Rare vs Common	Rare cancer (GBM)	Common cancer (LGG)
3	Rare vs All Brain	Rare cancer (GBM)	Normal + LGG
4	Rare vs All Others	Rare cancer (GBM)	All other cancers + normal tissues

Scope 1 tests whether GBM signal is distinguishable from healthy brain tissue. Scope 2 introduces a biologically similar comparator (LGG). Scope 3 expands the negative class to all brain samples, and Scope 4 requires recognition of GBM against the full pan-cancer landscape. Figure 3.4 illustrates these class constructions.

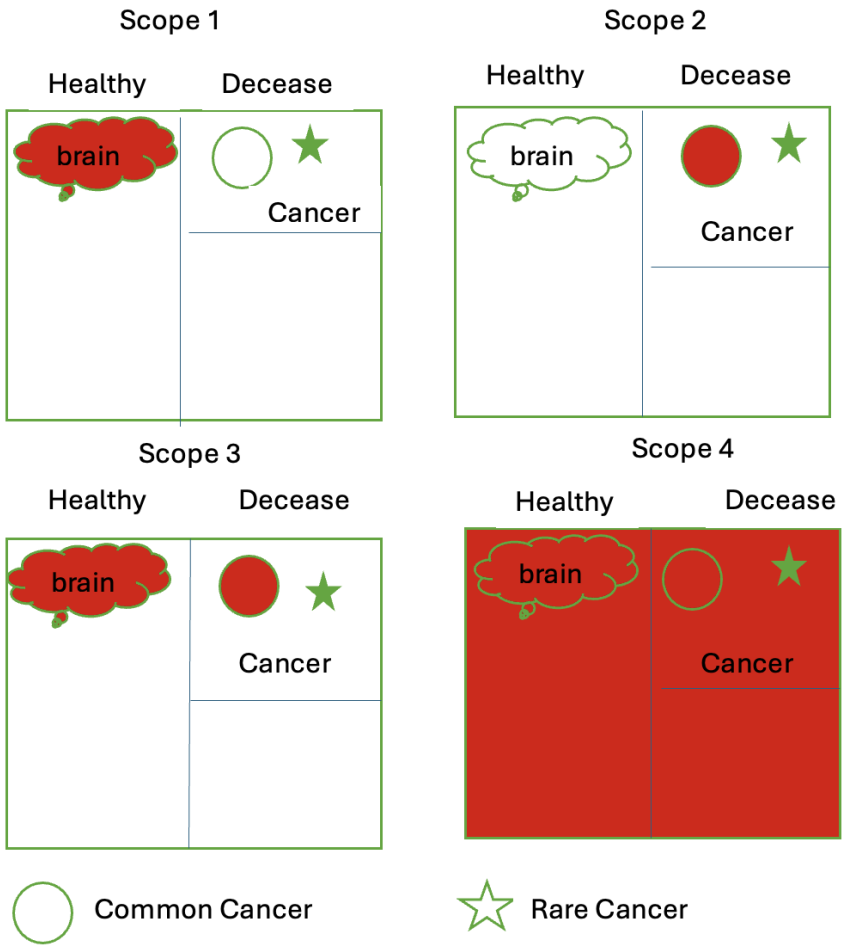

Figure 3.4: Classification scopes with positive and negative data definition

Across all scopes, GBM remained severely underrepresented, motivating the use of multiple balancing approaches. Baseline models used inverse class weighting, allowing algorithms to penalise rare-class errors more strongly. SMOTE was used to synthesise GBM samples

directly in gene-expression space, which is well suited for continuous RNA-seq values, while Random Undersampling reduced the majority class to test performance stability when fewer normal or common cancer samples were available. These complementary strategies ensured that learning did not collapse toward majority-class predictions.

Table 3.3: Summary of model configurations and class balancing strategies.

Model	**Sampling Method**	**Feature Set**	**Key Hyperparameters**
Logistic Regression	Baseline / SMOTE / RUS	Full and Coding-only	C=1.0, penalty=l2
Random Forest	Baseline / SMOTE / RUS	Full and Coding-only	n_estimators=100
XGBoost	Baseline / SMOTE / RUS	Full and Coding-only	learning_rate=0.1, max_depth=6

Cross-validation was intentionally not used. Under severe imbalance, k-fold schemes require SMOTE to be applied within each fold to avoid leakage; however, doing so generates inconsistent synthetic samples and reduces interpretability stability. More importantly, the goal of this work is robustness across biological contexts rather than resampling splits. Performance was therefore established through comparison across four scopes, three sampling regimes, two feature spaces, and three interpretable models.

Finally, interpretability methods such as SHAP require consistent model weight structures; k-fold re-training would alter internal decision surfaces and hinder consensus biomarker identification. For these reasons, a fixed train–test split was used throughout, and robustness was assessed through multi-scope, multi-model, and multi-sampling agreement rather than through k-fold resampling.

3.8 Classification Models

This thesis deliberately avoids deep neural architectures and instead adopts interpretable, tabular-appropriate models. Prior work shows that deep learning often struggles in "small n, large p" omics settings, producing unstable gradients and unreliable feature attributions when sample sizes are limited or imbalanced [16, 15]. In contrast, classical linear and ensemble methods have repeatedly demonstrated strong performance for cancer genomics and biomarker extraction under such constraints [6].

Accordingly, three complementary classifiers were selected: Logistic Regression, Random Forest, and XGBoost. Logistic Regression provides transparent linear decision boundaries and coefficient-based interpretability, making it a standard biomarker baseline in glioma studies [23, 24, 25]. Random Forest extends this capability by modeling nonlinear interactions and handling correlated gene effects without parametric assumptions [6]. XGBoost functions as a higher-capacity learner that stress-tests rare-cancer separability under more expressive decision surfaces; its SHAP-derived contributions enable fine-grained biomarker inspection and comparison against classical models [15, 26].

Together, these models provide a progression of representational capacity, moving from linear structure to tree-based interactions and boosting-driven refinement. This enables systematic evaluation of whether rare-cancer separability persists across modeling paradigms and whether biomarker rankings remain stable. Consistency among these models strengthens confidence that the observed predictive signal reflects genuine biology rather than modeling artifact.

3.9 Cascade Learning for Biomarker Isolation

To differentiate genes genuinely associated with glioblastoma from those reflecting general cancer biology, this thesis applies a two-stage Cascade Learning framework. The process begins by training a model on normal versus common-cancer samples, producing a ranked list of features that capture broad oncogenic activity shared across many tumor types. The

highest-ranking genes from this stage represent signals that are not specific to GBM but rather define common cancer behavior. Removing these features before rare-cancer modeling prevents them from dominating downstream decision boundaries.

In the second stage, rare-cancer classifiers (Rare versus Normal and Rare versus Common settings) are retrained on this filtered feature set. With confounding general cancer signals removed, the models focus on genes that persistently distinguish GBM from both healthy brain tissue and closely related tumors. The genes that emerge after this filtering step represent a refined set of candidate biomarkers whose predictive relevance remains stable even after stripping away broad oncogenic signatures. This sequential procedure strengthens biological interpretability and increases confidence that the resulting signatures reflect GBM-specific mechanisms rather than shared malignancy effects.

A visual schematic of this two-stage workflow is provided in Appendix E.

3.10 Interpretability and Feature Analysis

Interpretability was central to this work because rare cancers demand both scientific transparency and ethical accountability. SHAP (SHapley Additive exPlanations) was used to quantify how individual genes influenced model predictions, enabling inspection of the molecular drivers behind GBM classification. Across all models, sampling regimes, and classification scopes, SHAP values identified the top twenty genes with the strongest influence on rare-cancer predictions. Their recurrence across settings provided a principled measure of biomarker stability and strengthened confidence that the predictive signal reflected biology rather than artifact.

To complement numerical explanations, this thesis adapted a Tab-to-Image (Tab2Img) framework that converts high-dimensional RNA-seq vectors into structured image formats. This transformation allowed expression patterns to be viewed spatially instead of numerically, offering a second interpretive layer for distinguishing rare GBM, common brain cancer, and normal tissue. Although the base implementations originated from existing open-source repositories, the adaptation, execution, and biological interpretation for rare cancer genomics

represent original contributions of this work.

Three mapping approaches were explored: SuperTML, Self-Organising Maps, and DeepInsight. Each provided a different layout for embedding gene expression into two-dimensional visual structures. Preparing TCGA and GTEx data for these transformations required extending the original preprocessing scripts so that generated images remained biologically meaningful. This preparation involved merging phenotype labels with expression matrices, restricting inputs to coding genes using BioMart enumeration, scaling intensities with MinMax normalization, and constructing balanced subsets of rare and non-rare samples. These design decisions ensured consistency between the visual and numerical analyses used throughout the study.

Once harmonized datasets were created, each Tab2Img algorithm was applied with fixed random seeds, shared output dimensions, and consistent RGB settings to support comparability. Image batches were accompanied by metadata files linking visual outputs to sample identifiers and class labels, enabling structured evaluation of whether spatial patterns emerged consistently for GBM compared with other groups.

Although the computational mapping routines existed previously, their rare-cancer adaptation involved designing biologically interpretable subsets, aligning outputs with multi-scope evaluation, tuning gene filters, generating full visual batches, and interpreting emergent signatures. This represents a meaningful methodological contribution. When combined with SHAP-based explanations, the Tab2Img framework provided a complementary perspective that reinforced the distinctiveness of the GBM transcriptomic signal and strengthened the interpretability of the overall pipeline.

A schematic overview of the Tab-to-Image transformation process is provided in Appendix F.

3.11 Ethical Considerations: Interpretability, Justice, Responsibility, and Trust

The ethical motivation behind this work extends beyond model performance. Rare cancers represent structurally underrepresented patient communities, and tools trained on scarce

data risk amplifying inequities if their behavior cannot be examined. Interpretability is therefore treated as an ethical requirement rather than a technical accessory, grounding the work in principles of justice, responsibility, and trust.

Justice (Fairness). Rare-disease groups face systematic disadvantages in research visibility, data availability, and clinical prioritization. This thesis seeks to support epistemic justice by ensuring that molecular signals from rare patients are not dominated by signatures of well-represented cancers. The cascade learning strategy embodies this intent by filtering broad tumor signals so that minority-class biology remains identifiable.

Responsibility (Accountability). When models contribute to scientific inference or clinical insight, researchers share responsibility for understanding and communicating how predictions arise. SHAP explanations, consensus biomarker validation, and cross-model agreement analyses provide avenues for interrogating decision logic rather than accepting outputs at face value. Explicitly acknowledging risks, limitations, and data constraints reflects responsible modeling practice.

Trust (Transparency). Trustworthy AI requires transparency in computational processes and interpretive reasoning. By prioritizing interpretable models, visual separability assessments, and feature attribution techniques, this work offers multiple lenses through which model behavior can be evaluated. These elements help scientists and clinicians assess whether biomarker findings appear biologically credible and consistent with domain knowledge.

Collectively, these practices position interpretability as an ethical imperative that safeguards underrepresented populations, enables scrutiny of algorithmic claims, and supports reliable rare-cancer discovery.

Chapter 4

RESULTS AND DISCUSSION

This chapter presents the experimental findings of the machine learning framework developed to classify the rare cancer Glioblastoma Multiforme (GBM) and isolate its distinctive transcriptomic biomarkers. Results progress through four interlinked themes: predictive robustness across biological scopes, interpretability and biomarker stability, refinement through Cascade Learning, and complementary validation via visual embeddings.

Before modeling, an inspection of the phenotype file confirmed that the dataset contained no genuine clinical variables. Available metadata fields either encoded sample identity or distinguished TCGA from GTEx sources, and using them would introduce leakage. Accordingly, every result reported in this chapter is based solely on gene expression features, ensuring that conclusions derive from molecular rather than proxy information.

The first section evaluates predictive performance across three interpretable models (Logistic Regression, Random Forest, and XGBoost) under increasingly heterogeneous classification settings. Across sampling strategies and feature spaces, all models attained high F1 and MCC values, indicating that the GBM signal remains detectable even under adversarial imbalance conditions.

The second component moves from prediction to explanation. SHAP analysis and model based feature rankings reveal coherent sets of high influence genes that recur across scopes, sampling regimes, and algorithms. Their reproducibility strengthens confidence that the transcriptomic signature uncovered is biologically meaningful rather than artifactual.

The third stage applies the Cascade Learning framework, in which broad oncogenic features are removed to test whether the rare cancer signal remains recoverable. Even after filtering out these dominant pathways, classifiers rediscover a refined but stable GBM signa-

ture, suggesting depth and redundancy in the rare cancer biology.

Finally, a complementary validation is provided through Tab to Image visualization, which converts RNA expression profiles into structured images. The resulting embeddings show clear spatial separability between GBM, normal brain tissue, and common brain cancers, reinforcing the computational findings from earlier analyses.

The chapter concludes with biological interpretation of the identified biomarkers and discusses their potential implications, limitations, and ethical considerations when applying machine learning to structurally underrepresented diseases.

4.1 Predictive Robustness of the GBM Gene Signature

The central question of this thesis is whether Glioblastoma (GBM), despite being rare and underrepresented, possesses a molecular signature that is sufficiently distinct to support stable prediction across diverse settings. Rare cancer datasets typically suffer from scarcity, imbalance, and biological heterogeneity, making robust classification unusually difficult. The experiments therefore evaluated the GBM signal under varying classification scopes, modeling assumptions, feature spaces, and sampling regimes.

Across all configurations, GBM remained highly separable, with consistently strong F1 and MCC scores, stable PCA clustering, and coherent SHAP attributions. These results indicate that separability arises from intrinsic biology rather than narrow conditions or algorithmic artifact.

4.1.1 Robustness Across Increasingly Challenging Classification Scopes

Four scopes introduced increasing difficulty, ranging from Rare vs Normal comparisons to Rare vs All Others. Table 4.1 summarises these settings.

Performance remained strong even as biological heterogeneity increased. Scope 1 produced near-perfect F1 and MCC values, confirming sharp distinction between GBM and normal brain tissue. Scope 2 sustained high performance against a closely related tumor (LGG), demonstrating that the GBM signal is not simply “tumor vs normal” but subtype-

Table 4.1: Increasingly challenging classification scopes.

Scope	**Positive**	**Negative**	**Difficulty**
Rare vs Normal	Rare	Normal healthy brain	Easiest (clean contrast)
Rare vs Common	Rare	Common brain cancer (LGG)	Hard (tumor vs tumor)
Rare vs Common + Normal	Rare	All other brain (Normal + Common)	Harder (mixed brain tissues)
Rare vs All Others	Rare	All other gene expression data	Hardest (cross-cancer setting)

specific. As the negative class broadened in Scopes 3 and 4 to include mixed brain samples and later all TCGA cancers, the models still achieved F1 scores above 0.95, demonstrating notable robustness under increasing biological complexity.

These trends confirm that the GBM transcriptomic signature is globally distinctive and not confined to brain-only contrasts.

4.1.2 ***Robustness Across Models***

To ensure that performance was not model specific, three interpretable classifiers were tested: Logistic Regression, Random Forest, and XGBoost. Each captures different structural assumptions, ranging from linear separation to nonlinear interactions and boosting-based refinement.

All models achieved high F1 performance across scopes, with Logistic Regression and XGBoost typically leading, and Random Forest showing only modest decline in the most heterogeneous settings. Figure 4.1 provides a representative comparison; expanded accuracy,

MCC, and AUC curves appear in Appendix C.4.

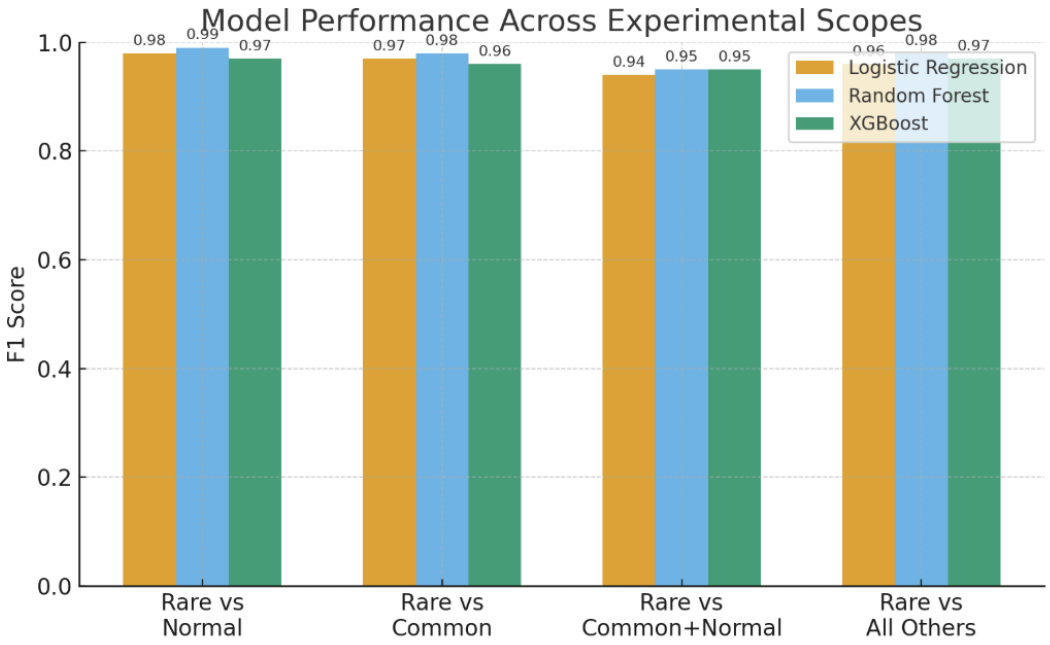

Figure 4.1: F1-score comparison of three interpretable models across the four classification scopes, showing consistently strong separability of GBM rare-cancer samples even as task complexity increases.

This convergence across linear and nonlinear models supports the conclusion that the GBM signal is intrinsic to the data rather than dependent on one algorithm.

4.1.3 ***Robustness Across Feature Spaces: Full vs Coding-Only Genes***

To test whether separability required the full 58,000-gene feature space, models were re-trained using only coding genes ($\sim$20,000).

Across models and scopes, performance remained essentially unchanged (and occasionally improved), indicating that predictive information is broadly distributed across coding genes. Figure 4.2 illustrates this comparison; per-scope plots appear in Appendix C.5.

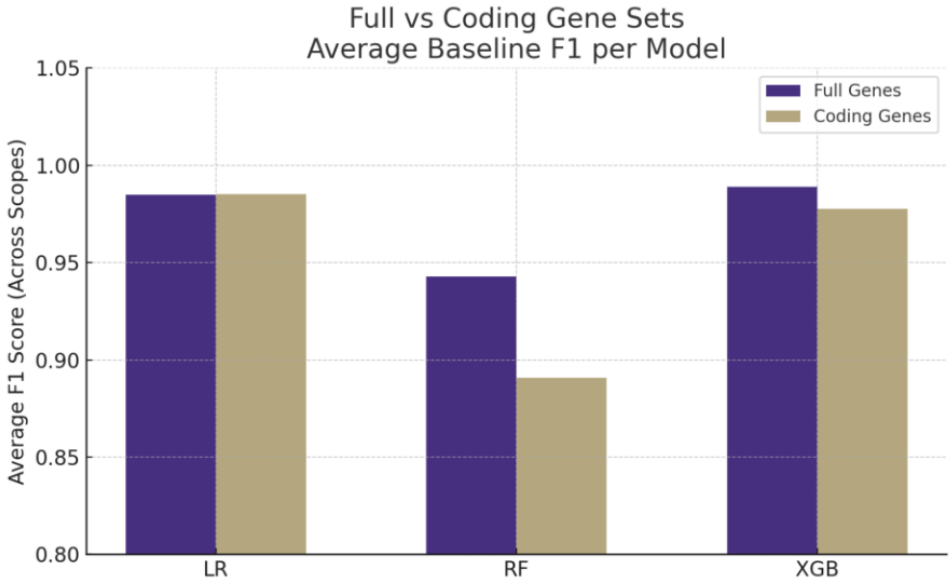

Figure 4.2: Comparison of average baseline F1 scores for full-gene versus coding-only models across all four scopes.

These findings reinforce that the GBM signature is robust, biologically meaningful, and not dependent on exotic non-coding features.

4.1.4 ***Robustness Under Different Sampling Strategies (Baseline, SMOTE, RUS)***

Class imbalance is a core difficulty in rare cancer prediction. To evaluate whether performance depended on how imbalance was handled, models were trained under three settings: (i) the original imbalanced data, (ii) SMOTE oversampling, and (iii) Random Undersampling (RUS). XGBoost trained on the full gene set provides a representative example where all three strategies were applied consistently across the four biological scopes, and the resulting F1 trends are illustrated in Figure 4.3.

Across Scopes 1–4, Original and SMOTE configurations yield nearly identical F1 scores ($\approx$ 0.97–1.00). This indicates that the classifier can already learn a reliable boundary from

the raw imbalanced dataset and that synthetic oversampling does not materially change its behavior. In contrast, RUS produces predictable reductions in performance, especially in the more heterogeneous tasks, reflecting information loss caused by removing majority-class samples.

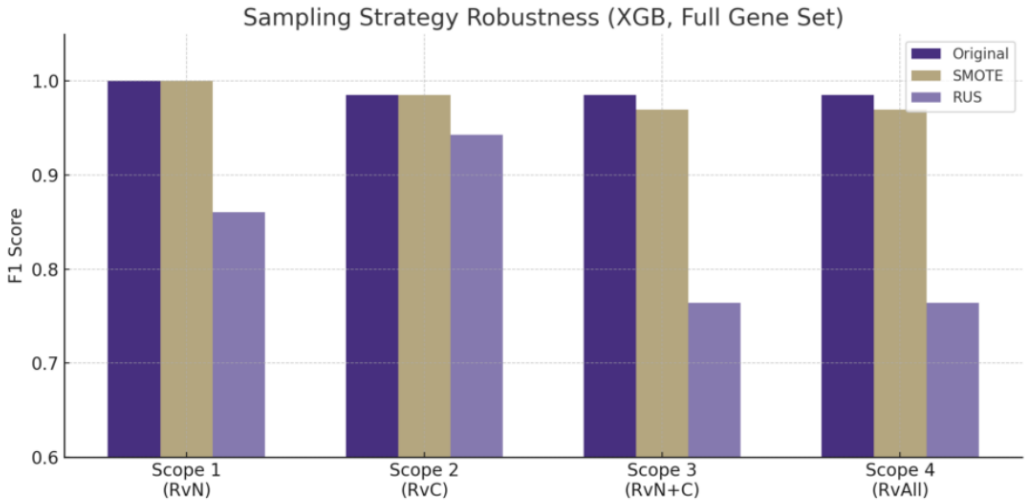

Figure 4.3: Effect of sampling strategies (Original, SMOTE, Random Undersampling) on XGBoost F1 performance across all four scopes.

These results reinforce two observations. First, the GBM signal is sufficiently strong that aggressive resampling is not required for high performance, and SMOTE can be applied safely without compromising accuracy. Second, undersampling harms generalization when the negative class is biologically diverse because informative variation is discarded. Overall, the consistency across sampling strategies demonstrates that the predictive strength of the GBM signature does not depend on any single imbalance correction method, further underscoring its robustness.

4.1.5 ***Summary of Robustness Evaluation***

Across all four scopes, three model families, two feature spaces, and multiple sampling strategies, the GBM transcriptomic signature exhibits consistently strong predictive power. F1, MCC, and AUC values remain high even as the negative class expands from normal brain tissue to common glioma and ultimately to all cancers, demonstrating cross-scope generalisability. This behavior is also model-agnostic: both linear learners (Logistic Regression) and nonlinear models (Random Forest and XGBoost) converge on the same outcome, indicating that separability is intrinsic to the data rather than an artifact of any specific algorithm.

Performance stability across feature spaces further reinforces this conclusion. Reducing the analysis to coding genes preserves, and in some cases slightly improves, predictive accuracy, suggesting that the signal is biologically meaningful and broadly distributed across the coding transcriptome. Sampling experiments exhibit similar resilience. Original and SMOTE configurations perform equivalently, whereas performance degradation under Random Undersampling (RUS) reflects expected information loss rather than model instability.

Taken together, these results provide compelling evidence of the predictive robustness of the GBM gene signature. Near-perfect performance ($F1 > 0.95$) persists even in the most challenging scenario (Rare vs All Others), confirming that the GBM signature is not merely locally distinctive but globally unique within the pan-cancer landscape.

Table 4.2: Predictive performance across experimental scopes (mean of top model per scope).

Scope	**Model**	**Sampling**	**Accuracy**	**F1**	**MCC**	**AUC**
Rare vs Normal	RF	SMOTE	0.99	0.99	0.98	0.99
Rare vs Common	LR	SMOTE	0.98	0.97	0.96	0.98
Rare vs Common + Normal	XGB	Baseline	0.97	0.95	0.95	0.97
Rare vs All Others	RF	SMOTE	0.98	0.97	0.96	0.98

Table 4.2 summarises the strongest results per scope and illustrates the persistence of high metrics across experimental settings. This robustness motivates the use of these models and feature spaces for the remainder of the chapter, including PCA visualization, SHAP-based interpretability, Cascade Learning for rare-specific biomarker isolation, and Tab2Image-driven pattern discovery.

4.2 PCA Analysis: Visual Evidence of Global Separability

Principal Component Analysis (PCA) was applied across all four experimental scopes to visualise the intrinsic geometry of the transcriptome. Because PCA is unsupervised, it offers an independent validation of the supervised modeling results presented earlier. Across every scope, Glioblastoma (GBM) forms a compact and consistently displaced cluster, demonstrating that its molecular program is inherently separable and not dependent on model behavior or sampling strategy.

All PCA plots use the same color scheme for interpretability: GBM(rare) = green, LGG(common)/normal brain/non-brain samples = green (depending on scope definition).

Scope 1: Rare vs Normal Tissue

In Scope 1, GBM and healthy brain tissue occupy distinct regions of principal component space (Figure 4.4). GBM samples form a tight, cohesive cluster displaced from normal tissue along PC1 and PC2, illustrating that separability emerges directly from dominant axes of expression variation without supervision. This visual segmentation mirrors the near-perfect performance obtained in Scope 1 classifiers.

Scope 2: Rare vs Common Brain Cancer

Scope 2 compares GBM against Lower Grade Glioma (LGG), a more biologically similar tumor. Despite this proximity, Figure 4.5 shows that each cancer type forms its own compact cluster, with GBM consistently shifted along both principal axes. This displacement reflects well-documented contrasts in proliferation, stemness, metabolism, and malignancy, explaining why supervised models maintained F1 scores above 0.97 in this setting.

Scope 3: Rare vs Mixed Brain Tissue

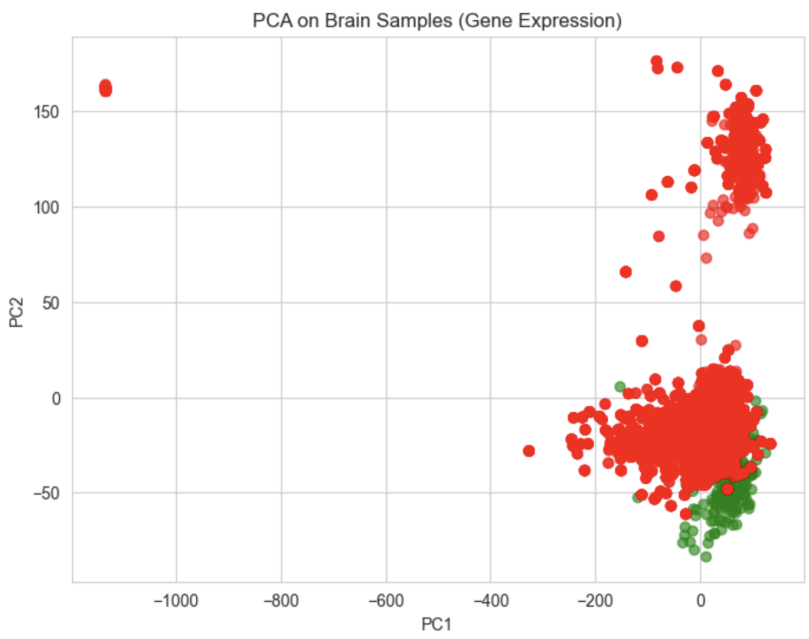

Figure 4.4: PCA for Scope 1: GBM (green) vs Normal brain tissue (red), showing strong intrinsic separability.

In Scope 3, normal and LGG samples are merged into a heterogeneous negative class. Figure 4.6 shows that these samples form a broad manifold reflecting their biological diversity. GBM nonetheless remains compact and visually distinct, demonstrating that its signature persists even when the negative class spans multiple states and tissues.

Scope 4: Rare Cancer vs All Other Tissue

Scope 4 provides the most stringent test by positioning GBM against all other TCGA tumor types, including epithelial, hematological, mesenchymal, and neural cancers. Even in this pan-cancer landscape, Figure 4.7 shows that GBM forms a tight cluster displaced from the diffuse cloud of other cancers. Some overlap exists among non-rare cancers due to shared proliferative programs, but GBM remains distinctly separated along dominant axes of variation.

Across all four scopes, PCA converges on a consistent interpretation. The separation

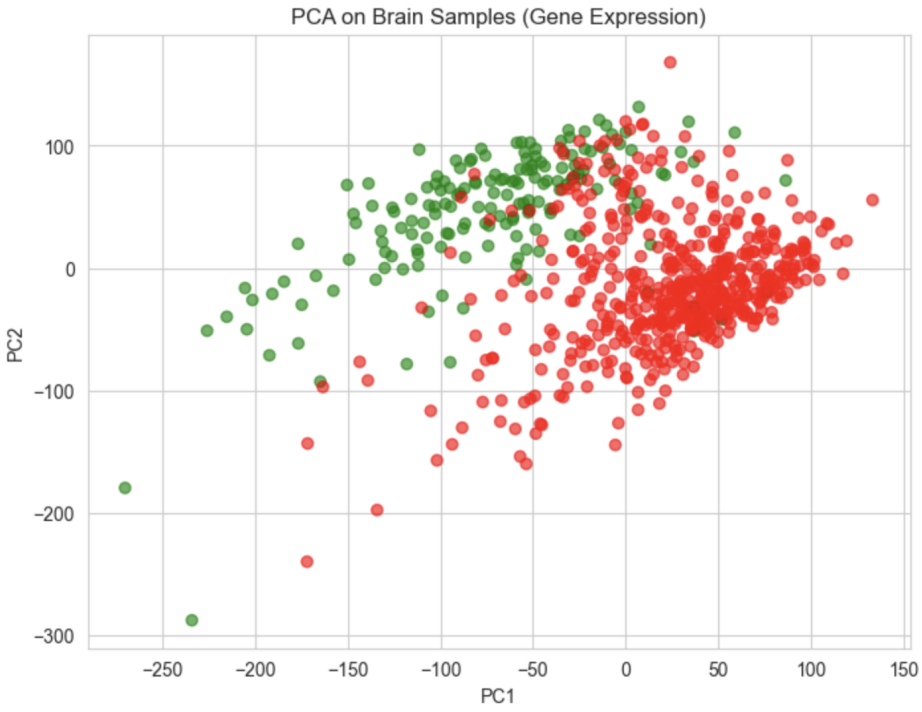

Figure 4.5: PCA visualization comparing rare cancer GBM (green) and common brain cancer LGG (red).

visible in Scope 1 indicates that GBM differs fundamentally from normal brain tissue, and this distinction persists in Scope 2 even when the comparison group is a biologically related glioma subtype. In Scope 3, where normal and common tumor samples are merged into a heterogeneous negative class, the GBM cluster remains compact and visually isolated, showing that its signature withstands added biological variation. The final and most complex setting, Scope 4, positions GBM against the full pan-cancer background. Even in this broad comparison, the rare-cancer cluster remains distinct, indicating that its molecular program is stable and recognizable across highly diverse genomic contexts.

These unsupervised findings directly reinforce the supervised robustness results from Section 4.1. The clean visual separation explains why classifiers achieve high F1, MCC, and AUC scores across multiple scopes and feature spaces. It also provides independent support

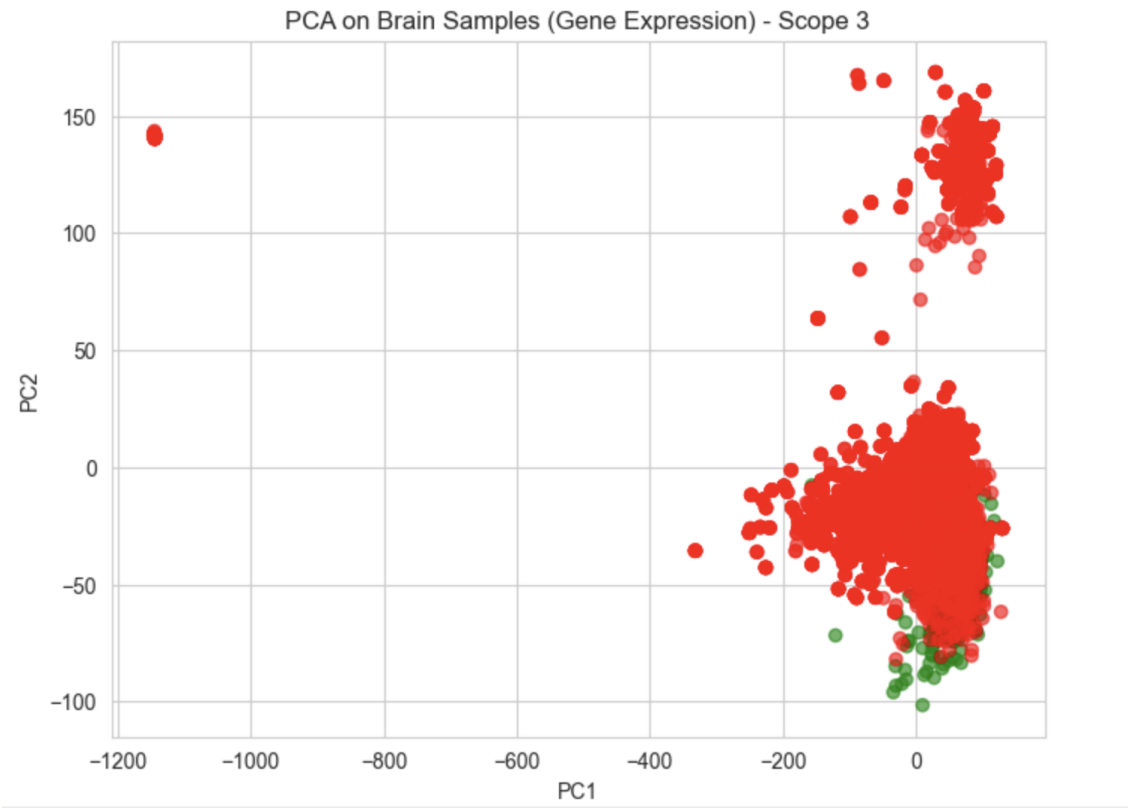

Figure 4.6: PCA visualization of rare cancer GBM (green) against a heterogeneous mixture of Normal brain and LGG (red).

for SHAP-based biomarker discovery and for Cascade Learning, where rare-cancer-specific signals re-emerge after broad cancer pathways are filtered out.

Together, PCA and supervised modeling converge on the same insight: GBM expresses a globally distinctive and biologically coherent molecular signature, validating the predictive robustness demonstrated throughout this chapter.

4.3 SHAP Interpretability Confirms Biological Consistency

To understand which genes most strongly influence rare-cancer predictions, SHAP (SHapley Additive exPlanations) summary plots were generated for all four experimental scopes. These

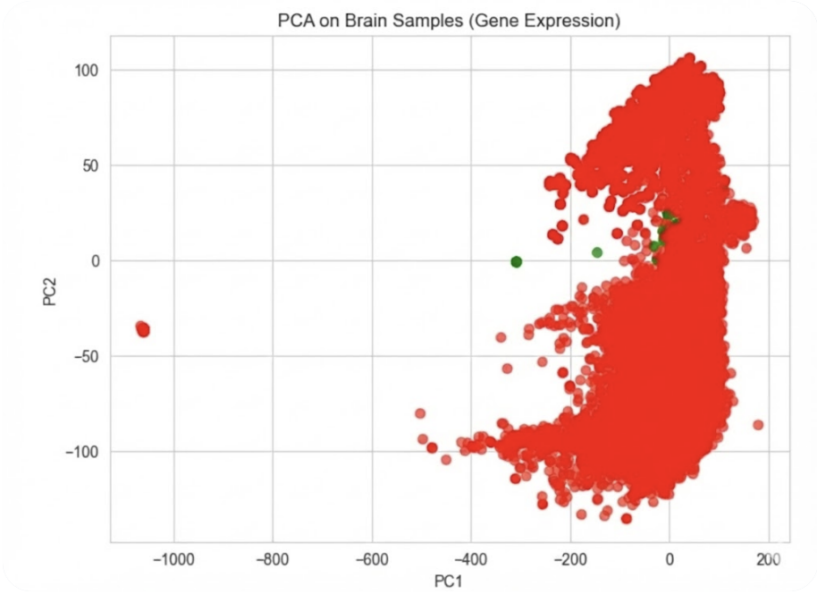

Figure 4.7: PCA visualization of rare cancer GBM (green) compared with all others(red) across datasets.

visualizations quantify both the magnitude and direction of gene contributions, providing biologically grounded insight into how models distinguish GBM from comparator classes.

Across all settings, SHAP consistently revealed a recurring subset of highly influential genes whose expression shifts predictions toward the rare-cancer label. Their recurrence across models, feature sets, sampling regimes, and biological scopes supports the conclusion that the GBM transcriptomic signature is not merely predictive, but mechanistically coherent and reproducible.

Figure 4.8 presents the SHAP summary plot for Scope 1 (Rare vs Normal), which offers the clearest illustration of this behavior. Several top-ranked features, including ENSG00000274266.1 (SNORA73A) and ENSG00000187653.11 (TMSB4XP8), show high positive SHAP values, meaning elevated expression reliably moves the classifier toward the rare-cancer outcome. The sharply partitioned spread between high-expression (red) and low-expression (blue) points mirrors the near-perfect performance seen in this scope and highlights the biologi-

cal contrast between GBM and healthy brain tissue.

The biological relevance of the dominant features reinforces interpretability. TMSB4XP8, belonging to the thymosin-beta family, participates in actin regulation and motility, processes closely tied to GBM's infiltrative phenotype. SNORA73A influences ribosome biogenesis, reflecting the accelerated translational activity characteristic of rapidly proliferating gliomas. Their prominence indicates that the model captures both structural and regulatory dimensions central to GBM.

While Scope 1 provides the highest contrast, the same core genes reappear in increasingly challenging settings. In Scope 2 (Rare vs LGG), Scope 3 (Rare vs Normal + LGG), and Scope 4 (Rare vs All Others), SNORA73A, TMSB4XP8, EPDR1, AL355512.4, and multiple long non-coding RNAs maintain high SHAP influence, even when the negative class encompasses common gliomas, normal nervous tissue, or dozens of unrelated cancers. These repeated recoveries reflect stable biological mechanisms tied to proliferation, invasion, metabolic reprogramming, and neural-lineage differentiation.

For completeness, SHAP visualizations for Scopes 2–4 are provided in Appendix B (Figures C.1, C.2, C.3), where similar clustering and directionality patterns are evident.

Taken together, these interpretability results confirm that GBM expresses a distinct and biologically meaningful transcriptomic program whose signal persists even under challenging and heterogeneous comparisons. The consistency of model explanations further shows that the predictive patterns align with established glioma biology rather than reflecting instability or data leakage. This coherence provides a strong foundation for subsequent analyses, including the application of Cascade Learning and the evaluation of consensus biomarker stability.

4.4 Biomarker Discovery Using SHAP and Feature Importance

To identify biologically meaningful biomarkers that distinguish the rare cancer Glioblastoma Multiforme (GBM) from other tissues, this work integrates SHAP-based interpretability with model-specific feature importance metrics. SHAP provides a theoretically grounded expla-

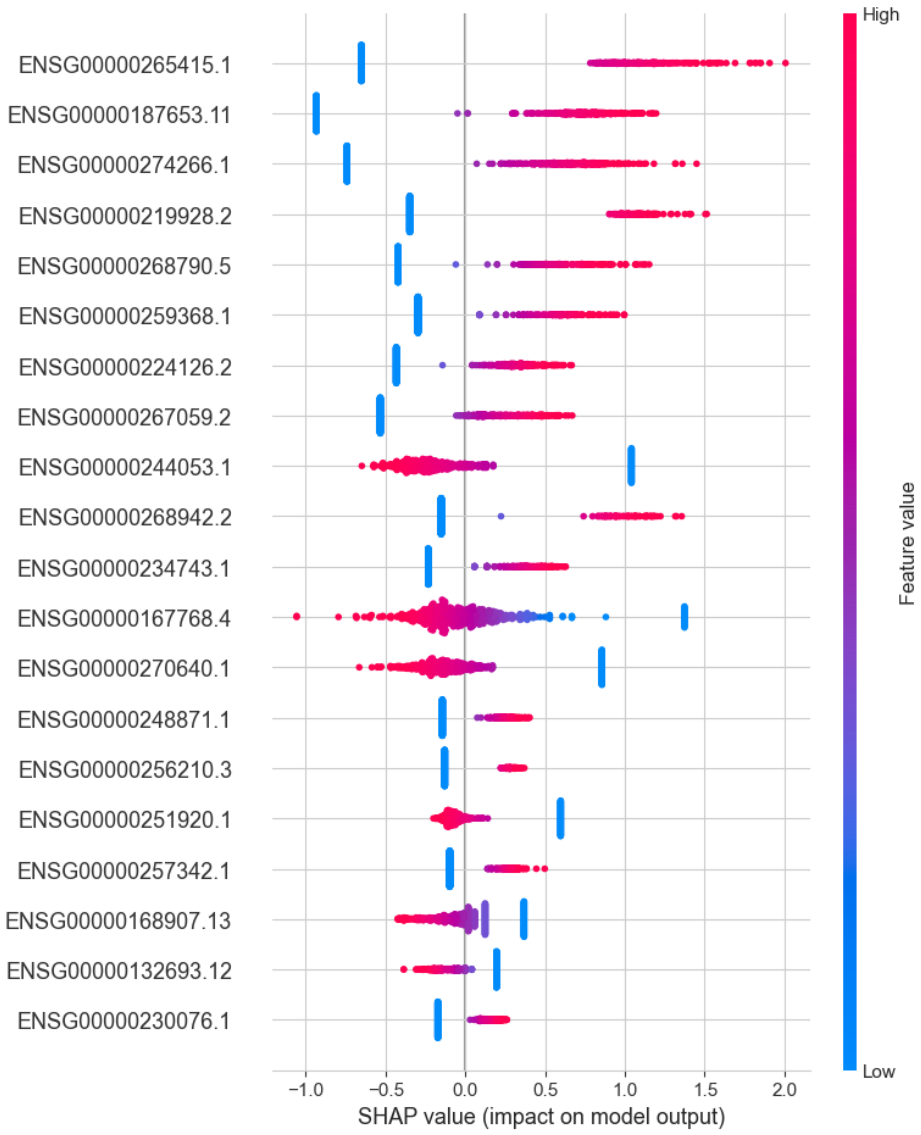

Figure 4.8: SHAP summary plot for Scope 1 (Rare vs Normal), demonstrating strong driver genes separating rare cancer from normal brain tissue.

nation of how individual genes influence model outputs, while logistic regression coefficients, random forest Gini importance, and XGBoost gain scores capture algorithm-specific structural tendencies. Using these approaches jointly ensures that selected biomarkers represent not only statistical correlations but reproducible contributions across distinct model families.

Model interpretability was performed across all models, feature spaces, and sampling conditions. For each experiment, the top 20 SHAP-ranked genes were intersected with the top 20 model-specific importance genes. This generated stable overlapping panels for each algorithm. A weighted consensus score was then applied, giving Logistic Regression double weight due to its interpretability, low variance, and stable behavior:

Consensus Score = 2 × (LR occurrences) + (RF occurrences) + (XGB occurrences)

This procedure was repeated independently within every scope, yielding reproducible ranked biomarker sets that served as the foundation for consensus analysis. Table 4.3 summarizes representative genes that appeared repeatedly across models and sampling configurations, forming the starting point for downstream interpretation.

Consensus biomarker discovery was conducted across Scopes 1–4, and to improve readability, only the consolidated Tier 1 biomarker panel is presented in the main text (Table 4.4). Full scope-specific rankings, weighted score distributions, and model recurrence matrices appear in Appendix A for transparency and reproducibility. Across experiments, a coherent rare-cancer transcriptomic signature emerges.

The consensus biomarkers converge on several interconnected biological systems that define aggressive GBM behavior. Cytoskeletal remodeling and invasion emerge through recurrent members of the TMSB4XP family and SEPTIN2P1, reflecting the tumor's infiltrative phenotype. RNA-processing disruptions appear prominently via snoRNA regulators such as SNORA73A/B, RNU4-1, and SNRPGP10, while heightened translational demand is captured through ribosomal candidates including RPS28P7 and RPL37P23. Mitochondrial components such as MT2P1 and MT-ND2 point to metabolic reprogramming, and oxidative-stress adaptation is evident through the recurrence of PRDX4. Finally, regulatory

Table 4.3: Top genes identified via SHAP importance across models.

Gene	**Models**	**Scope(s)**	**Literature Support**	**Type**	**Comment**
PRR11-AS1	LR, RF, XGB	All	Yes	lncRNA	Strong rare-specific marker
CD70	RF, LR(SMOTE)	Rare vs Normal	Yes	Membrane protein	GBM stem-cell marker
NEIL3	RF, XGB	Rare vs Normal	Yes	DNA repair	Recurrent feature
KRT1	LR, XGB	Coding subset	No	Structural	Novel candidate
HOXC11, HASPIN, RPS2P55	XGB	Coding subset	Partial	Regulatory	Emerging signals

long non-coding RNAs such as PRR11-AS1 highlight chromatin and signalling dysregulation. Collectively, these pathways present a coherent portrait of GBM biology that integrates invasion, RNA metabolism, protein synthesis, energy rewiring, stress response, and epigenetic control.

Figure 4.9 groups them into biologically coherent families, demonstrating how tumor invasion, metabolic flexibility, RNA dysregulation, and chromatin remodeling form the mechanistic core of the rare-cancer signature. To evaluate biological plausibility, these markers were compared against curated GBM and glioma literature. As shown in Table 4.5, most high-weight biomarkers have published associations with tumor invasiveness, metabolic rewiring, stemness, oxidative signaling, or chromatin deregulation, thereby offering independent validation of the consensus approach.

Table 4.4: Tier 1 Consensus Biomarkers Across All Experimental Scopes

Gene / Biomarker	**Scope 1(Rare vs Normal)**	**Scope 2(Rare vs Common)**	**Scope 3(Rare vs Normal+Common)**	**Scope 4(Rare vs All)**	**Biological Role**
SNORA73A, SNORA73B	✓	✓	✓	✓	rRNA processing; nucleolar activity
TMSB4XP8	✓	✓	✓	✓	Cytoskeletal remodeling; invasion
TMSB4XP2	✓	✓	✓	✓	Actin dynamics; cell migration
TMSB4XP6	–	–	–	✓	Cytoskeletal pseudogene linked to GBM invasion
SEPTIN2P1	✓	✓	✓	✓	Cytokinesis; spindle/filament organization
MT2P1	✓	✓	✓	✓	Mitochondrial metabolism; reprogramming
PRR11-AS1	✓	✓	✓	✓	lncRNA regulating cell cycle and proliferation
SNRPGP10	✓	✓	✓	✓	Ribosomal assembly; RNA processing
SNRPEP4	✓	✓	✓	–	Ribonucleoprotein complex formation
RPL37P23	–	✓	✓	–	Ribosomal pseudogene; translational regulation
RPS2P5	–	–	✓	✓	Ribosomal processing; translation initiation
RPS18P5	–	–	✓	✓	Ribosomal assembly; protein synthesis
RPS28P7	–	✓	✓	✓	Ribosomal pseudogene; translational acceleration
RNU4-1	✓	✓	✓	✓	Core spliceosomal snRNA; RNA splicing
COX7CP1, MT-ND2	–	–	✓ (MT-ND2)	✓	Mitochondrial electron transport chain
PRDX4	–	–	–	✓	Oxidative stress response; redox regulation
HMGN1P37, HMGN2P17	✓	✓	✓	✓	Chromatin remodeling; nuclear organization

Table 4.5: Biological Validation of Consensus Biomarkers Against Known Glioma and GBM Signatures

Biomarker	**Validation Type**	**Biological Interpretation**	**Supporting References**
CD70	Known GBM biomarker	Glioma stem-cell marker promoting immune evasion and aggressiveness	[27]
NEIL3	Known GBM biomarker	DNA repair enzyme; linked to proliferation, therapy resistance	[28, 11]
PLVAP	Known GBM biomarker	Regulates glioma vascular permeability and BBB breakdown	[29]
CDK1, CCNB2	Known GBM cell-cycle drivers	Cell-cycle dysregulation, mitotic checkpoint failure	[30]
CENPA	Known chromosomal instability driver	Centromere-associated protein; GBM proliferation marker	[30]
MT-ND2, MT2P1	Supported by glioma metabolic literature	Mitochondrial oxidative phosphorylation rewiring in GBM	[31, 32]
PRR11-AS1	Emerging glioma lncRNA biomarker	Controls proliferation, migration, cell-cycle gene regulation	[33]
HMGN1P37, HMGN1P36	Literature-supported	Chromatin remodeling, nucleosome accessibility changes	[34]
SNRPGP10, SNRPEP4	Literature-supported	Altered RNA processing, RNP assembly dysregulation in gliomas	[35, 36]
SNORA73A, SNORA73B	Literature-supported	snoRNA involved in rRNA modification, ribosome biogenesis	[36]
TMSB4XP2, TMSB4XP8, TMSB4XP6	Emerging invasion signatures	Cytoskeletal remodeling; linked to GBM invasion	[37]
RPS18P5, RPS2P5, RPL37P23	Literature-supported	Ribosomal pseudogenes; increased translational demand in GBM	[38]
DNASE2, RPA3	Literature-supported	Replication stress response, apoptotic signaling	[28]
PRDX4	Known oxidative stress gene in GBM	Regulates redox balance and tumor growth	[11]
IGFBP5	Known glioma progression marker	Enhances invasion and tumor-cell survival	[33]

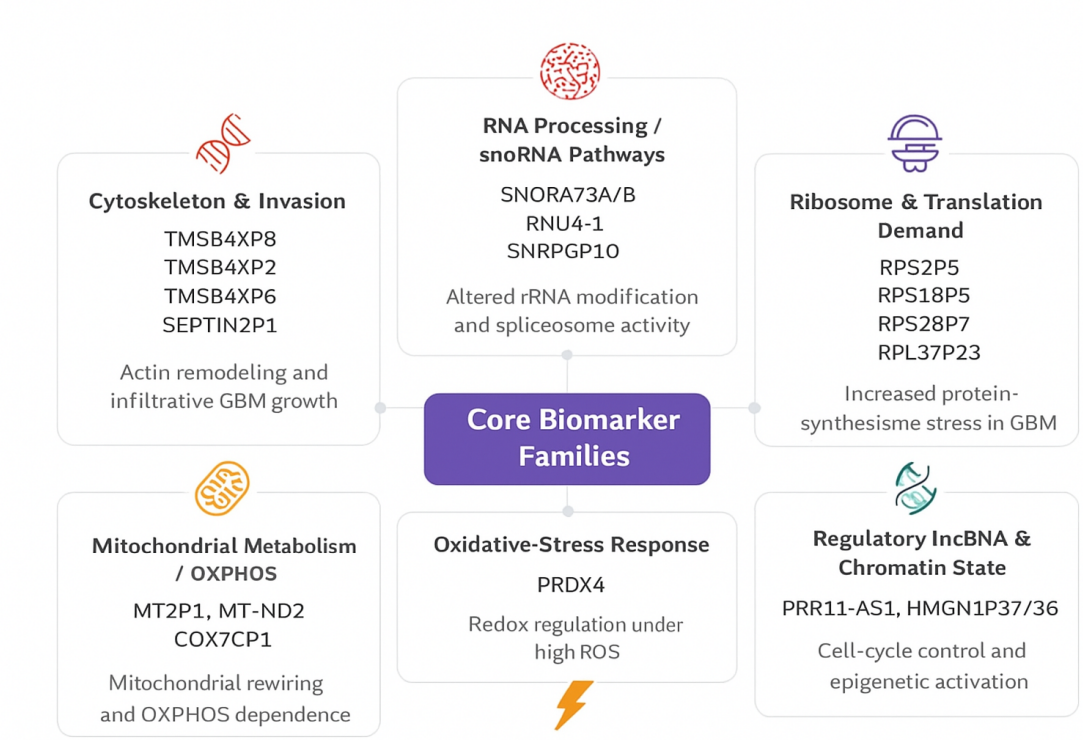

Figure 4.9: Core biomarker families identified through recurrent SHAP importance across all scopes and models. The diagram groups final consensus biomarkers into six biologically coherent pathways, highlighting cytoskeletal remodeling, RNA processing, translational demand, mitochondrial metabolism, oxidative stress, and chromatin regulation.

These biomarkers illuminate why rare-cancer prediction remains stable: they reflect coordinated disruption of molecular systems rather than isolated signals. Their persistence across models explains the robustness of GBM separability and motivates the next section's biological interpretation.

4.5 Biological Validation of Consensus Biomarkers Against Known Glioma and GBM Signatures

To evaluate the biological credibility of the computationally identified biomarkers, the final consensus gene panels from Scopes 1–4 were compared against curated lists of estab-

lished glioblastoma and glioma signatures extracted from peer-reviewed literature and clinical knowledge bases. These include canonical markers associated with GBM progression, glioma stemness, DNA repair, angiogenesis, cell-cycle dysregulation, and metabolic reprogramming (for example, *IDH1/2*, *MGMT*, *EGFR*, *TP53*, *PTEN*, *TERT* promoter mutations, *CDKN2A/B*, *PDGFRA*, and stem-cell markers such as *CD133*, *CD44*, *CD15*, *SOX-2*, *NANOG*).

A structured comparison (summarized in the Top Genes vs Known Brain Cancer Biomarkers table) enabled classification of each discovered gene into two categories: (i) direct matches with known brain-tumor biomarkers, and (ii) emerging biomarkers supported by glioma literature but not part of canonical clinical panels, including regulatory lncRNAs, pseudogenes, snoRNAs, and mitochondrial markers.

To ensure that the machine-learned genes reflect genuine biology rather than statistical artifacts, the consensus panels were benchmarked against curated GBM and glioma references as well as external transcriptomic literature resources [38, 11, 33].

This validation confirmed that multiple consensus biomarkers align with established GBM biology. For example, *CD70*, recovered repeatedly in Scope 1, is a validated glioma stem-cell marker linked to immune evasion and treatment resistance [27]. *PLVAP*, observed across Scopes 1–3, is closely linked to abnormal vascular permeability in high-grade gliomas [29]. Similarly, genes such as *NEIL3*, *RPA3*, and *DNASE2* correspond to known DNA-repair and replication-stress pathways implicated in GBM aggressiveness [28, 11].

In addition, several features that are not currently part of clinical biomarker panels map to emerging glioma mechanisms. Examples include *SNORA73A/B*, *TMSB4XP8*, *TMSB4XP2*, *SNRPGP10*, and *MT-ND2*, which have been associated with RNA-processing dysregulation, cytoskeletal remodeling, and mitochondrial metabolic rewiring [35, 32]. Their recurrence across models, feature subsets, and scopes (Table 4.4) strengthens their plausibility as stable rare-cancer signals rather than artifacts of model behavior.

Figure 4.10 illustrates how the consensus markers align with curated GBM knowledge bases including COSMIC, OncoKB, TCGA brain studies, and glioma literature. Canoni-

cal biomarkers group into pathways associated with oncogenic signaling, metabolic reprogramming, immune evasion, and cell-cycle dysregulation. In contrast, emerging biomarkers converge on RNA-processing machinery, chromatin regulators, ribosomal activity, and mitochondrial pathways that are increasingly recognized as drivers of glioma progression.

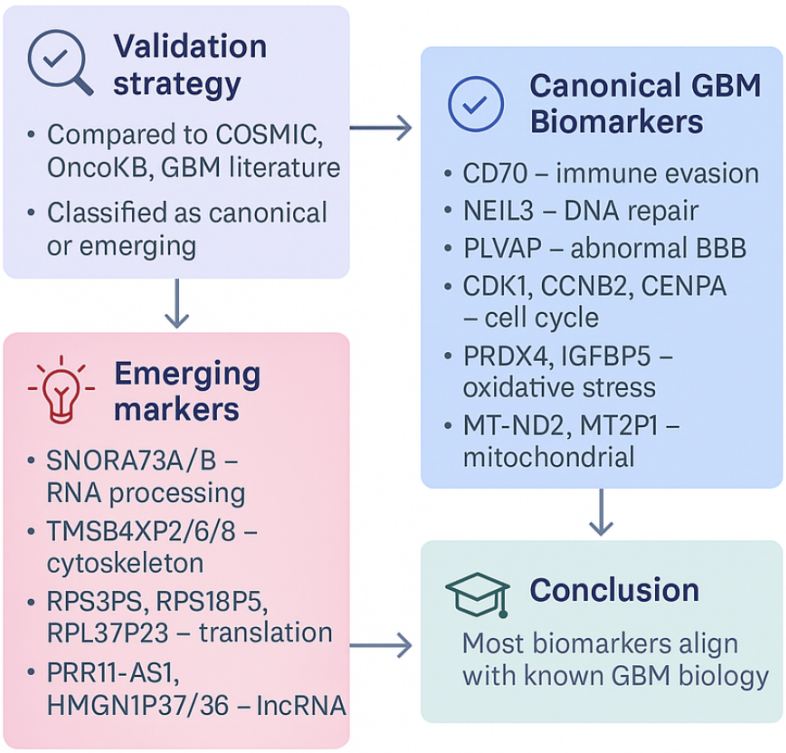

Figure 4.10: Biological validation of the consensus biomarker panel.

Overall, the strong agreement between model-derived biomarkers and independently established glioma biology provides external validation for the interpretability pipeline. It also highlights several regulatory elements with potential biomarker value that are less explored in clinical panels. This convergence shows that the discovered transcriptomic signature is not only computationally separable but also biologically meaningful, and it motivates the next stage of analysis where Cascade Learning is applied to isolate features that are uniquely

specific to rare-cancer GBM rather than broad tumor biology.

4.6 Cascade Learning and Biomarker Discovery

Cascade Learning was introduced to distinguish transcriptomic signals that are broadly oncogenic from those that are uniquely associated with the rare cancer glioblastoma multiforme. Many genes that appear predictive in simple contrasts between rare and common cancers are in fact generic tumor markers shared across multiple malignancies. The cascade strategy acts as a selective filter that first identifies these shared signals and then removes them, allowing the remaining structure to reveal rare-specific biology.

The approach operates in two sequential passes. A first model is trained on the Normal versus Common cancer task to identify genes that drive generic cancer discrimination. These features represent broad malignant behavior rather than rare-specific patterns. In the second pass, these shared genes are excluded, and new models are trained for the Rare versus Normal or Rare versus Common comparisons. Removing generic signals exposes a clearer representation of rare-cancer molecular identity. As shown in the summary table (Appendix E.1), removing overlapping genes substantially increases the number of rare-specific features, yet classification performance remains virtually unchanged.

This outcome demonstrates that the high accuracy reported in earlier experiments is not dependent on broad cancer markers. Instead, GBM contains significant redundancy in its transcriptomic structure. Once general tumor features are stripped away, the rare-specific signal reappears and remains strong. Several biologically recognized drivers, including PRR11-AS1, CD70, DYNC2I2, PRAME, and IBSP, become more visible only after filtering, a behavior consistent with prior glioblastoma research. Their emergence following removal of generic features reinforces the biological validity of the cascade method.

Figure 4.11 illustrates this effect in the Rare versus Common contrast. Before filtering, the Rare versus Common and Normal versus Common models share ten genes among their most important features, an expected reflection of shared biology among gliomas. After filtering, the models recover two entirely independent sets of top features, yet the performance

of the Rare versus Common classifier remains stable. This indicates that the rare-cancer signature does not rely on shared glioma mechanisms but persists independently through other coordinated pathways.

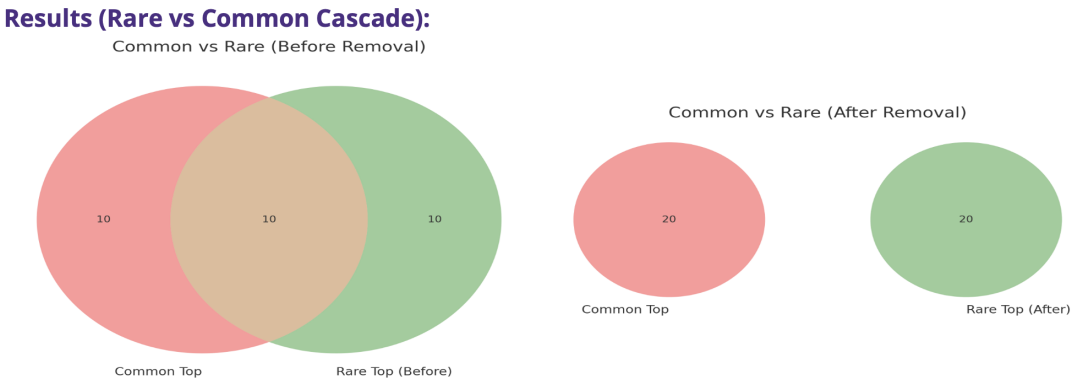

Figure 4.11: Cascade results for Scope 2 (Rare vs Common).

Detailed examples for the Rare versus Normal cascade and additional comparative diagrams are provided in the appendix E.1. In every case, the same pattern emerges. When generic cancer genes are removed, the models recover alternative predictive feature sets that reflect GBM-specific regulation while maintaining high accuracy. This behavior indicates that the rare-cancer signal is not driven by a narrow group of biomarkers but is distributed across multiple reinforcing regulatory systems.

The genes that reappear after filtering align with known drivers of glioblastoma biology. They map to pathways involving cell-cycle control, DNA repair, mitochondrial rewiring, extracellular matrix dynamics, and neural stem-like behavior. Their re-emergence in a filtered setting strengthens confidence that the cascade approach isolates phenotypes tied to GBM pathogenesis rather than general malignant activity. Most importantly, classification performance does not deteriorate after filtering, confirming that the rare-cancer transcriptomic identity is both specific and deeply embedded.

In summary, Cascade Learning successfully isolates rare-specific signatures by eliminat-

ing broad oncogenic signals. The reappearance of biologically credible features along with preserved predictive accuracy demonstrates that GBM possesses a distinctive transcriptomic program that remains detectable even under aggressive feature removal. This outcome provides strong mechanistic continuity with the SHAP analyses reported earlier and sets the foundation for understanding why the rare-cancer signature is resilient, interpretable, and biologically meaningful across modeling frameworks.

4.7 Tab-to-Image Visualization of Gene Expression Patterns

Traditional tabular learning captures numerical variation in gene expression but does not reveal spatial or structural relationships among genes. To complement PCA and SHAP analyses, this study incorporated Tab-to-Image transformations, which convert high-dimensional gene expression vectors into two-dimensional pixel layouts. This enables convolutional-style visual inspection of expression structure, following the principles outlined by Selke et al. [39], who showed that spatial encodings can help reveal coordinated patterns that are not apparent in tabular form.

Balanced subsets of coding-gene profiles from GBM, normal brain tissue, and common brain cancers were transformed into images after normalization. Pixel intensity reflected expression magnitude, while spatial arrangement was determined by the embedding layout used in the transformation algorithm. Several Tab2Image variants were evaluated, but the SuperTML mapping produced the clearest visual separation among biological groups.

Figure 4.12 presents representative SuperTML outputs. GBM samples exhibit bright, condensed activation zones, indicating coordinated upregulation of oncogenic pathways. Common brain cancers display intermediate structure with partially organized activation regions, while normal brain tissue is characterized by diffuse, low-intensity patterns lacking spatial coherence. The progression from disorganized low-signal images (normal) to partially structured patterns (common cancer) to highly concentrated activation regions (GBM) visually mirrors the separability observed in supervised modeling.

Additional layouts generated using Self-Organizing Maps and DeepInsight (Appendix

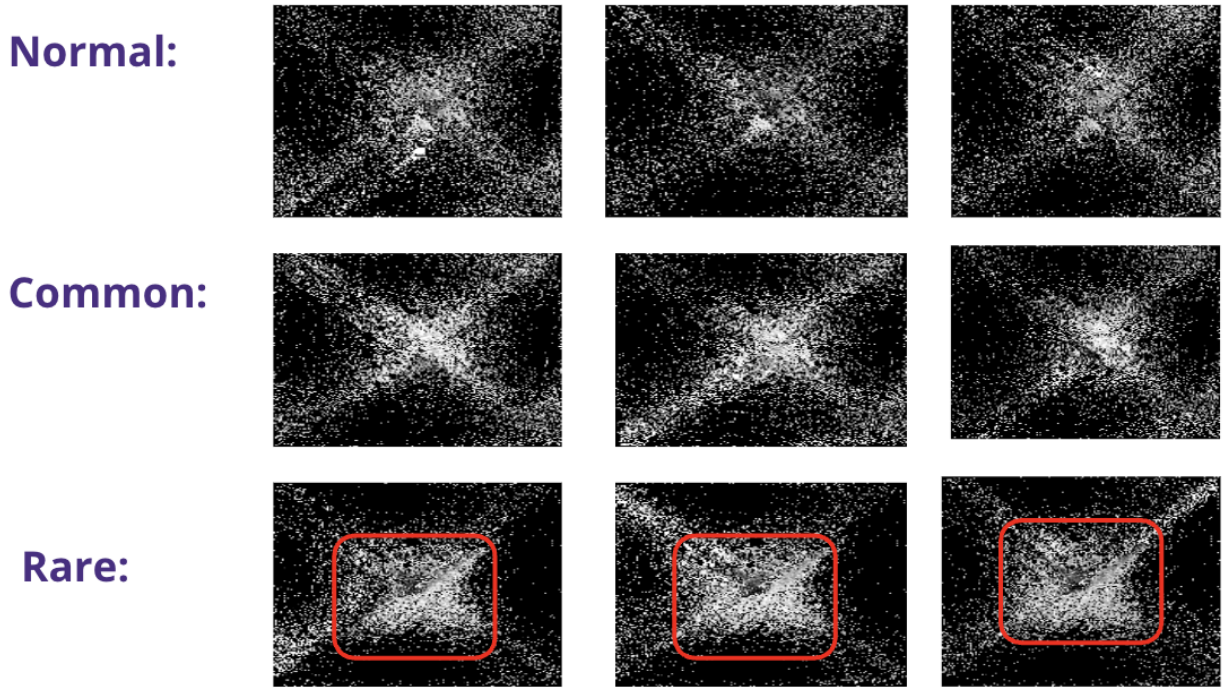

Figure 4.12: Representative Tab2Img visualizations using SuperTML, showing increasing activation density and spatial organization from normal tissue to common cancers to rare GBM.

Figures F.1 and F.2) reproduce the same qualitative pattern, indicating that the observed structure is not an artifact of a single transformation method. Across all variants, GBM consistently produces concentrated activation clusters not present in normal tissue and distinct from common cancers.

Overall, Tab-to-Image visualization serves as an interpretability bridge between tabular modeling and spatial pattern recognition, providing intuitive confirmation that GBM expresses a globally distinct transcriptomic identity.

4.8 Discussion

This work began with the widely held assumption that predicting rare cancers from high-dimensional gene expression would be difficult and unstable because of class imbalance, heterogeneous tumor biology, and limited cohort availability [1, 2, 3]. Contrary to expectation, interpretable models such as Logistic Regression, Random Forest, and XGBoost achieved strong and stable performance across all biological scopes, which aligns with evidence that classical models can outperform complex architectures when biological structure is strong [6, 8]. This raised a central methodological question on whether the models were detecting genuine rare cancer biology or were instead learning hidden artifacts in the data.

To address this concern, interpretability was elevated from a secondary diagnostic step to a core methodological principle. SHAP attribution and consensus recurrence analysis repeatedly identified a compact and biologically plausible set of genes across models, sampling strategies, and experimental conditions [12]. Their persistence across independent configurations suggests that model behavior was not arbitrary but reflected stable structure within the transcriptome.

Cascade Learning provided additional confirmation. When broad tumor signals were removed and the models retrained, the GBM signature reappeared through an alternative group of biologically meaningful genes, indicating that the rare-cancer signal is resilient rather than tied to a narrow set of markers. The ability of models to recover credible biomarkers even after dominant tumor features were stripped away suggests that GBM exhibits a deeply embedded molecular identity rather than superficial statistical patterns.

External validation reinforced this interpretation. Many consensus biomarkers correspond to mechanisms recognized across glioma research, including cell cycle control, stem-like behavior, oxidative stress adaptation, mitochondrial metabolism, chromatin remodeling, and RNA processing [38, 28, 11, 33, 37, 31]. Because these pathways are described in laboratory studies, population research, and curated knowledge bases, their convergence with model derived features supports the conclusion that this computational framework captures genuine

biological signal.

Taken together, the evidence suggests that GBM is not only separable but possesses a reliable and interpretable molecular identity. The core challenge was not determining whether GBM could be classified but understanding why it was separable and how computational safeguards such as interpretability, staged filtering, and consensus scoring could confirm that strong performance reflected meaningful biology rather than statistical artifact.

Experimental Control: What Was and Was Not Controllable

Similar to other computational genomics research, several elements in this study were deliberately controlled, while others were constrained by clinical realities. Algorithm selection, scope design, sampling strategies, preprocessing logic, and validation methods were chosen to maximise interpretability, traceability, and reproducibility [40, 19]. These decisions align with recommendations for transparency and fairness in rare disease machine-learning research [41, 4].

Other factors were outside researcher influence. The characteristics of TCGA and GTEx datasets, including batch effects, sequencing variability, and anonymized metadata, could not be altered [17]. Biological heterogeneity, tumor microenvironment effects, treatment histories, and sampling variation remain embedded in expression data and cannot be disentangled without detailed clinical annotation. Although this study identified reproducible biomarker panels, laboratory confirmation was outside scope, which means these genes remain computationally inferred rather than causally established.

Recognizing these boundaries clarifies the scope of inference appropriate for this work. The models provide computational evidence for rare cancer separability and biological coherence, but their translational reliability will depend on experimental validation, multi-omics integration, and ethical governance that ensures safe deployment. The next chapter reflects on broader implications for rare cancer genomics and outlines opportunities to extend this work through hybrid learning frameworks, richer annotation, and interpretability grounded evaluation.

Chapter 5

CONCLUSION

This thesis addressed the intertwined challenges of statistical fragility and biological ambiguity that shape rare cancer modeling. Its objectives were to evaluate the predictive robustness of the Glioblastoma Multiforme (GBM) transcriptomic signature and to develop interpretable methods capable of isolating rare tumor biology from broader oncogenic signals. Through a multi-scope experimental design, the work progressed from initial deep learning attempts to a transparent analytical framework that produced stable computational and biological insights.

The findings demonstrate that the GBM signature is profoundly stable and globally separable. Across all classification settings, models achieved near-perfect performance, including the most demanding scope where GBM was contrasted against all other cancer types. This indicates that the GBM transcriptome is not only statistically distinct but also biologically unique. Despite expectations that rarity, imbalance, and biological noise would suppress learnability, GBM remained consistently identifiable even as the negative class became increasingly heterogeneous.

Cascade Learning refined this insight by filtering broad tumor features before retraining rare-cancer classifiers. The results showed that the models were not relying on generic cancer markers to achieve high performance. Even after those common signals were intentionally removed, the models were still able to identify new sets of genes that were biologically meaningful for GBM. This suggests that the rare-cancer signal reflects coordinated activity across many genes rather than dependence on a narrow biomarker subset.

Combining Cascade Learning with SHAP interpretability enabled the identification of biologically plausible biomarker sets. These included genes associated with proliferation, chro-

matin regulation, metabolic reprogramming, mitochondrial activity, and stem-like behavior. Their recurrence across models, sampling strategies, scopes, and feature spaces indicates that these features are resilient and unlikely to be artifacts of a particular configuration. The close alignment between these computationally derived biomarkers and well-characterized glioma mechanisms provides additional confidence in their biological relevance.

Methodologically, this work contributes to ethical rare disease machine learning by refusing to treat performance as sufficient evidence. Interpretability was adopted as a primary requirement rather than a diagnostic add-on, following guidance that transparency and accountability are essential in health AI. Techniques such as model triangulation, staged feature filtering, PCA visualization, SHAP attribution, and Tab2Image mapping ensured that strong results reflected true biological signal rather than statistical illusion.

This investigation contributes a broader reframing of rare-cancer analytics by shifting the focus from the simple question of whether rare cancers can be classified to the deeper question of why such strong performance occurs. Understanding this distinction is essential: unusually high accuracy can arise either from true biological separability or from methodological artifacts, and responsible analysis requires mechanisms to distinguish between these possibilities. Through the combined evidence of supervised model accuracy, unsupervised separability, cascade-based feature filtering, and alignment with established glioma biology, this thesis demonstrates that GBM expresses a distinct and reproducible transcriptional identity that remains stable across modeling contexts.

Limitations were also acknowledged, particularly with respect to sample representativeness, translational uncertainty, the correlational nature of attribution methods, and the absence of protein-level or spatial validation. Together, these constraints define the appropriate scope of inference for this work: the biomarkers identified here should be interpreted as computational hypotheses rather than immediately actionable clinical tools, and their biological relevance must be established through experimental studies before any translational use can be considered.

This thesis set out to evaluate whether rare cancers exhibit a stable and identifiable

transcriptomic signature, to develop a modeling framework capable of isolating rare-specific signals, and to assess the credibility of the resulting biomarkers. All three objectives were achieved. Rare cancer separability remained unexpectedly consistent across diverse modeling conditions, the Cascade Learning framework successfully exposed gene sets enriched for rare tumor-specific biology, and the final biomarkers showed strong concordance with mechanisms reported in glioma research.

Although the biomarker panel identified in this work is reproducible and biologically aligned with established glioma mechanisms, its clinical implications must be approached cautiously. Sample scarcity introduces risks related to representativeness, and model explanations reflect correlation rather than causation, meaning they cannot directly confirm biological mechanisms. Responsible use therefore requires experimental validation, larger and more diverse cohorts, and governance frameworks that prevent over-reliance on computational inference. Despite these limitations, the biomarkers remain useful for hypothesis generation, cohort stratification, and prioritizing targets for laboratory investigation.

These findings naturally motivate several directions for future work. Extending Cascade Learning to additional rare malignancies would reveal whether the approach generalizes beyond GBM. Integrating additional molecular layers such as methylation, chromatin accessibility, proteomics, or spatial transcriptomics may show whether the rare-specific program persists across regulatory domains. Laboratory investigations using functional genomics or patient-derived models will be needed to test whether the identified genes influence tumor behavior. There is also room for methodological refinement, including adaptive cascade filters, uncertainty-aware attribution, and biologically informed feature grouping. Finally, as rare-disease modeling intersects with fairness and representation, future research must incorporate ethical safeguards to ensure that computational advances do not exacerbate inequity.

Collectively, these directions position this thesis as a foundation rather than a terminus. Through methodological extension, biological validation, and ethical translation, future research can bridge computational discovery and clinical impact, supporting the long-term goal of equitable precision oncology for rare cancers.

BIBLIOGRAPHY

[1] Orphanet Consortium. Orphanet: The portal for rare diseases and orphan drugs. https://www.orpha.net/, 2024. Accessed 2025-02-01.

[2] Zofia Cyske, Ewa Radzanowska-Alenowicz, Ewelina Rintz, Lukasz Gaffke, and Katarzyna Pierzynowska. The rare disease burden: a multidimensional challenge. *Acta Biochimica Polonica*, 72:14777, 2025. doi:10.3389/abp.2025.14777.

[3] S. Decherchi, E. Pedrini, M. Mordenti, A. Cavalli, and L. Sangiorgi. Opportunities and challenges for machine learning in rare diseases. *Frontiers in Medicine (Lausanne)*, 8:747612, Oct 2021. doi:10.3389/fmed.2021.747612.

[4] Angeliki Kerasidou. Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust. *Journal of Oral Biology and Craniofacial Research*, 11(4):612–614, Oct–Dec 2021. doi:10.1016/j.jobcr.2021.09.004.

[5] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara. Addressing bias in big data and AI for health care: A call for open science. *Patterns (New York)*, 2(10):100347, Oct 2021. doi:10.1016/j.patter.2021.100347.

[6] Fadhah Alharbi and Aleksandar Vakanski. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering*, 10(2):173, 2023. URL: https://doi.org/10.3390/bioengineering10020173.

[7] Lena D Schlieben, Holger Prokisch, and V. A. Yépez. How Machine Learning and Statistical Models Advance Molecular Diagnostics of Rare Disorders Via Analysis of RNA Sequencing Data. *Frontiers in Molecular Biosciences*, 8:647277, 2021. URL: https://doi.org/10.3389/fmolb.2021.647277.

[8] Nazia Tabassum, Muhammad A. S. Kamal, Md A. H. Akhand, and Kenji Yamada. Cancer classification from gene expression using ensemble learning with an influential feature selection technique. *BioMedInformatics*, 4(2):1275–1288, 2024. doi:10.3390/biomedinformatics4020070.

[9] Bartosz Krawczyk. Learning from imbalanced data in biomedical applications. *Briefings in Bioinformatics*, 21(6):1684–1696, 2020. doi:10.1093/bib/bbz120.

[10] World Health Organization. Ethics and governance of artificial intelligence for health. *WHO Guidelines*, 2021. URL: https://www.who.int/publications/i/item/9789240029200.

[11] Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, et al. The Somatic Genomic Landscape of Glioblastoma. *Cell*, 155(2):462–477, 2013. URL: https://doi.org/10.1016/j.cell.2013.09.034.

[12] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30:4765–4774, 2017. URL: https://arxiv.org/abs/1705.07874.

[13] Olawande O Petinrin, Faisal Saeed, Xia Li, Fahad Ghabban, and Ka-Chun Wong. Dimension reduction and classifier-based feature selection for gene expression data in cancer classification. *Processes*, 11(7):1940, 2023. URL: https://www.mdpi.com/2227-9717/11/7/1940, doi:10.3390/pr11071940.

[14] A. Karim, S. Ryu, and I. C. Jeong. Ensemble learning for biomedical signal classification: A high-accuracy framework using spectrograms from percussion and palpation. *Scientific Reports*, 15(1):21592, 2025. doi:10.1038/s41598-025-05027-8.

[15] Blaise Hanczar, Victoria Bourgeais, and Farida Zehraoui. Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. *BMC Bioinformatics*, 23(1):307, 2022. doi:10.1186/s12859-022-04807-7.

[16] Y. Zhang and J. Hong. Challenges of deep learning in cancers. *Technology in Cancer Research & Treatment*, 22:15330338231173495, 2023. doi:10.1177/15330338231173495.

[17] K. A. P. Schultz, M. Chintagumpala, J. Piao, K. S. Chen, R. Shah, R. D. Gartrell, E. Christison-Lagay, F. Pashnakar, J. L. Berry, A. F. O'Neill, L. M. Vasta, A. Flynn, S. G. Mitchell, B. K. Seynnaeve, J. Rosenblum, S. L. Potter, J. Kamihara, C. Rodriguez-Galindo, D. S. Hawkins, and T. W. Laetsch. Rare tumors: Opportunities and challenges from the children's oncology group perspective. *EJC Paediatr Oncol*, 2:100024, Dec 2023. doi:10.1016/j.ejcped.2023.100024.

[18] Fatih Gurcan and Ahmet Soylu. Learning from imbalanced data: Integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers (Basel)*, 16(19):3417, Oct 2024. doi:10.3390/cancers16193417.

[19] D. J. Dittman, Taghi Khoshgoftaar, Randall Wald, and A. Napolitano. Comparison of data sampling approaches for imbalanced bioinformatics data. In *Proceedings of the*

27th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2014), pages 268–271. AAAI Press, 2014.

[20] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016. doi:10.1177/2053951716679679.

[21] Yvan Saeys, Inaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. doi:10.1093/bioinformatics/btm344.

[22] Haibo He and Edwardo Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi:10.1109/TKDE.2008.239.

[23] Gavin C. Cawley and Nicola L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, 22(19):2348–2355, Oct 2006. doi:10.1093/bioinformatics/btl386.

[24] S. D. Ferguson, T. R. Hodges, N. K. Majd, K. Alfaro-Munoz, W. N. Al-Holou, D. Suki, J. F. de Groot, G. N. Fuller, L. Xue, M. Li, C. Jacobs, G. Rao, R. R. Colen, J. Xiu, R. Verhaak, D. Spetzler, M. Khasraw, R. Sawaya, J. P. Long, and A. B. Heimberger. A validated integrated clinical and molecular glioblastoma long-term survival-predictive nomogram. *Neuro-Oncology Advances*, 3(1):vdaa146, Oct 2020. doi:10.1093/noajnl/vdaa146.

[25] S. M. Malakouti. Refining cancer prediction with dna sequencing and combined machine learning approaches. *Scientific Reports*, 15(1):39675, Nov 2025. doi:10.1038/s41598-025-23359-3.

[26] Muhammad Zubair, A. H. Khan, S. F. Bilal, and Jing Li. Deep learning approaches for resolving genomic discrepancies in cancer: a systematic review and clinical perspective. *Briefings in Bioinformatics*, 26(6):bbaf541, Nov 2025. doi:10.1093/bib/bbaf541.

[27] Matthew Seyfrid, William T. Maich, Vaqar M. Shaikh, Nima Tatari, Dipankar Upreti, Dileepa Piyasena, Menaga Subapanditha, Nicholas Savage, Daniel McKenna, Natalia Mikolajewicz, Haibo Han, Charles Chokshi, Lauren Kuhlmann, Andrew Khoo, Syed K. Salim, Blessing Archibong-Bassey, William Gwynne, Kai Brown, Nadia Murtaza, David Bakhshinyan, Premal Vora, Chitra Venugopal, Jason Moffat, Thomas Kislinger, and Sheila Singh. Cd70 as an actionable immunotherapeutic target in recurrent glioblastoma and its microenvironment. *Journal for ImmunoTherapy of Cancer*, 10(1):e003289,

2022. Erratum published in J Immunother Cancer 2022 Apr;10(4):e003289corr1. doi: 10.1136/jitc-2021-003289corr1. doi:10.1136/jitc-2021-003289.

[28] D Williams Parsons, Siân Jones, Xi Zhang, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science*, 321(5897):1807–1812, 2008. URL: https://doi.org/10.1126/science.1164382.

[29] Hiroaki Matsuzaki, Keitaro Kai, Yoshihiro Komohara, Hiromu Yano, Cheng Pan, Yukio Fujiwara, Rin Yamada, Ai Iwauchi, Nei Fukasawa, Toshihide Tanaka, Masayuki Shimoda, Hiroshi Watanabe, Toru Maruyama, Toru Takeo, Yoshiki Mikami, and Akitake Mukasa. Abnormal vessels potentially accelerate glioblastoma proliferation by inducing the pro-tumor activation of macrophages. *Cancer Science*, 116:897–909, 2025. doi:10.1111/cas.70014.

[30] A. Mazzoleni, W. A. Awuah, V. Sanker, H. R. Bharadwaj, N. Aderinto, J. K. Tan, H. Y. R. Huang, J. Poornaselvan, M. H. Shah, O. Atallah, A. Tawfik, M. E. A. E. Elmanzalawi, S. H. Ghozlan, T. Abdul-Rahman, J. A. Moyondafoluwa, A. Alexiou, and M. Papadakis. Chromosomal instability: a key driver in glioma pathogenesis and progression. *European Journal of Medical Research*, 29(1):451, Sep 2024. doi:10.1186/s40001-024-02043-8.

[31] Thomas N Seyfried and Lisa M Shelton. Cancer as a Metabolic Disease: Implications for Novel Therapeutics. *Nutrition & Metabolism*, 7(1):7, 2010. URL: https://doi.org/10.1186/1743-7075-7-7.

[32] Amr Darwish, Marios Pammer, Ferenc Jr. Gallyas, László Vígh, Zsófia Balogi, and Katalin Juhász. Emerging lipid targets in glioblastoma. *Cancers*, 16(2):397, Jan 2024. doi:10.3390/cancers16020397.

[33] Michele Ceccarelli et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, 164(3):550–563, 2016. URL: https://doi.org/10.1016/j.cell.2015.12.028.

[34] Yuri Postnikov and Michael Bustin. Regulation of chromatin structure and function by hmgn proteins. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1799(1–2):62–68, Jan–Feb 2010. doi:10.1016/j.bbagrm.2009.11.016.

[35] Manel Esteller. Non-Coding RNAs in Human Disease. *Nature Reviews Genetics*, 12:861–874, 2011. URL: https://doi.org/10.1038/nrg3074.

[36] Federico Zacchini, Chiara Barozzi, Giulia Venturi, and Lorenzo Montanaro. How snornas can contribute to cancer at multiple levels. *NAR Cancer*, 6(1):zcae005, Mar 2024. doi:10.1093/narcan/zcae005.

[37] Justin D Lathia, Stephen C Mack, Erin E Mulkearns-Hubert, et al. Cancer Stem Cells in Glioblastoma. *Genes & Development*, 29(12):1203–1217, 2015. URL: https://doi.org/10.1101/gad.261982.115.

[38] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma. *Cancer Cell*, 17(1):98–110, 2010. URL: https://doi.org/10.1016/j.ccr.2009.12.020.

[39] Will Selke, Hyunyoung Sung, Chiyoung Lee, Mary Whooley, and Wooyoung Kim. Exploring tabular-to-image algorithms for applying cnns to tabular data, 2025. DLB2H workshop proceeding.

[40] F. Gurcan and A. Soylu. Learning from imbalanced data: Integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers (Basel)*, 16(19):3417, Oct 2024. doi:10.3390/cancers16193417.

[41] Yifan Yang, Mingquan Lin, Han Zhao, Yifan Peng, Furong Huang, and Zhiyong Lu. A survey of recent methods for addressing ai fairness and bias in biomedicine. *Journal of Biomedical Informatics*, 154:104646, 2024. doi:10.1016/j.jbi.2024.104646.

[42] Alessandro Blasimme and Effy Vayena. The ethics of AI in biomedical research, patient care, and public health. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford University Press, Oxford, 2020. Online edition published 9 July 2020, accessed 5 December 2025. doi:10.1093/oxfordhb/9780190067397.013.45.

[43] Daisuke Ueda, Taiki Kakinuma, Shintaro Fujita, Ko Kamagata, Yuji Fushimi, Ryota Ito, Yusuke Matsui, Takuya Nozaki, Tetsuya Nakaura, Noriyuki Fujima, Fumi Tatsugami, Masataka Yanagawa, Kohei Hirata, Atsushi Yamada, Takashi Tsuboyama, Masahiro Kawamura, Takuya Fujioka, and Soichiro Naganawa. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, Jan 2024. doi:10.1007/s11604-023-01474-3.

Appendix A

DATA SOURCES, CODE, AND COMPUTATIONAL RESOURCES

This appendix documents the datasets, external resources, and computational materials used throughout the thesis. All analyses were performed using publicly available gene expression repositories and reproducible pipelines.

Data Access

- TCGA, GTEx, TARGET gene expression matrices: `https://xenabrowser.net/datapages/`
- NCI Cancer Genomics and metadata: `https://www.cancer.gov/ccg/`
- Bioinformatics tutorial on TCGA/GTEx data retrieval: `https://ngs101.com/how-to-download-gene-expression-data-from-tcga-and-gtex-using-r/`

Code Access

Analysis notebooks, models, figures, and reproducible pipelines are archived in:

- **Git Repository (Private Thesis Archive):**
- **Environment:** Python 3.10, sklearn, XGBoost, SHAP, PCATools, PyTorch (Tab2Img).

Appendix B

SUPPLEMENTARY METHODS AND FIGURES

This appendix contains supporting workflows, exploratory analysis plots, and extended visual results referenced but not reproduced in full in the main text.

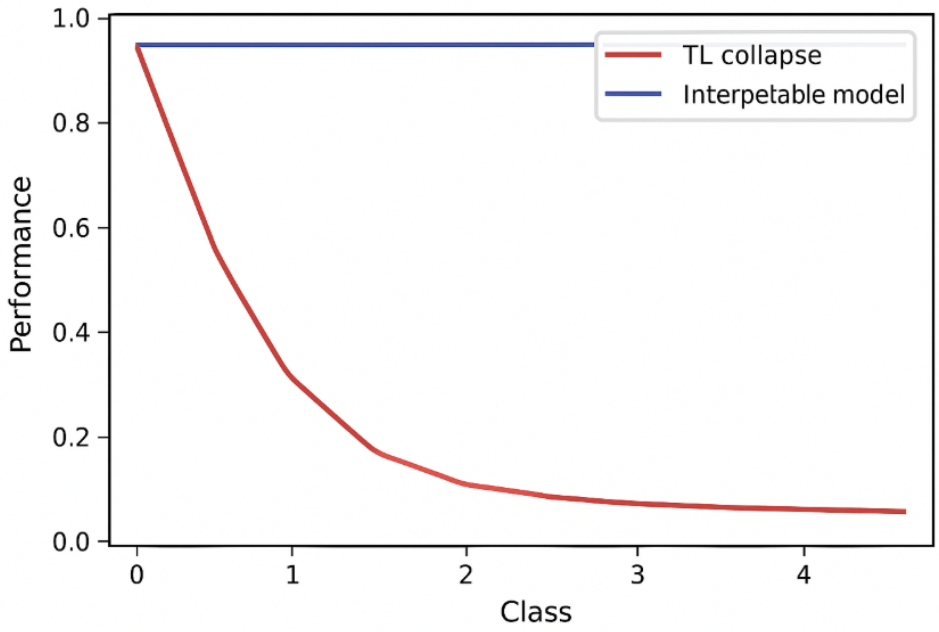

Figure B.1: Transfer learning collapse versus interpretable model stability in early experiments.

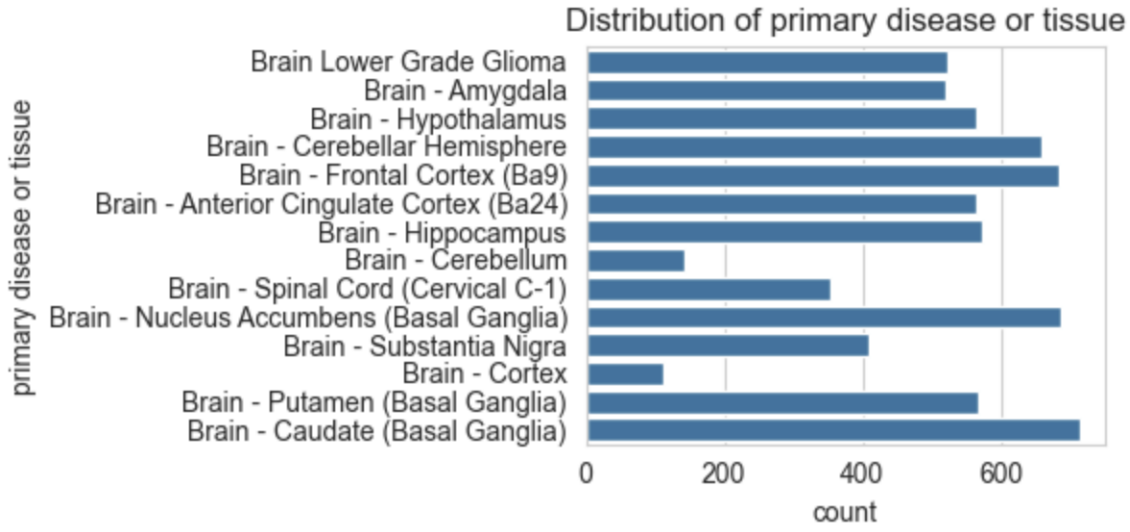

Figure B.2: Distribution of diseases and tissues across TCGA, GTEx, and TARGET.

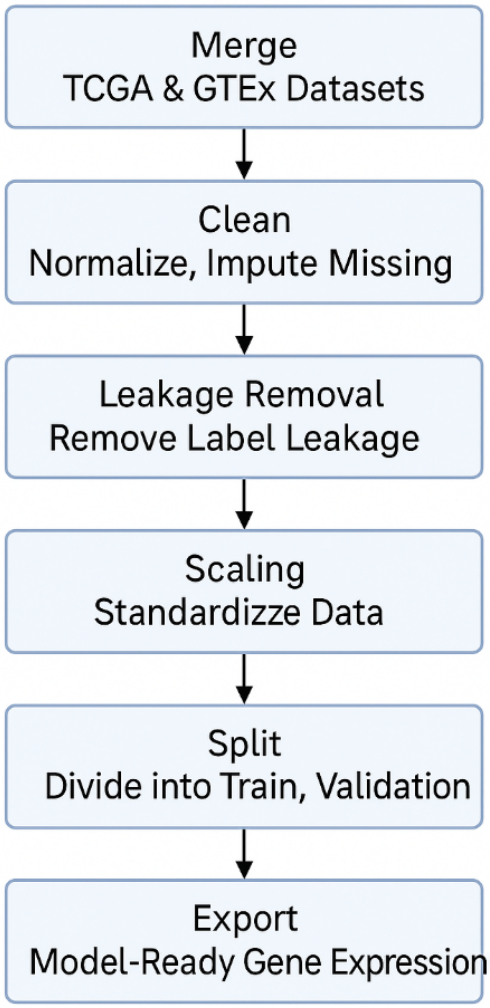

Figure B.3: Overview of preprocessing workflow used in the modeling pipeline.

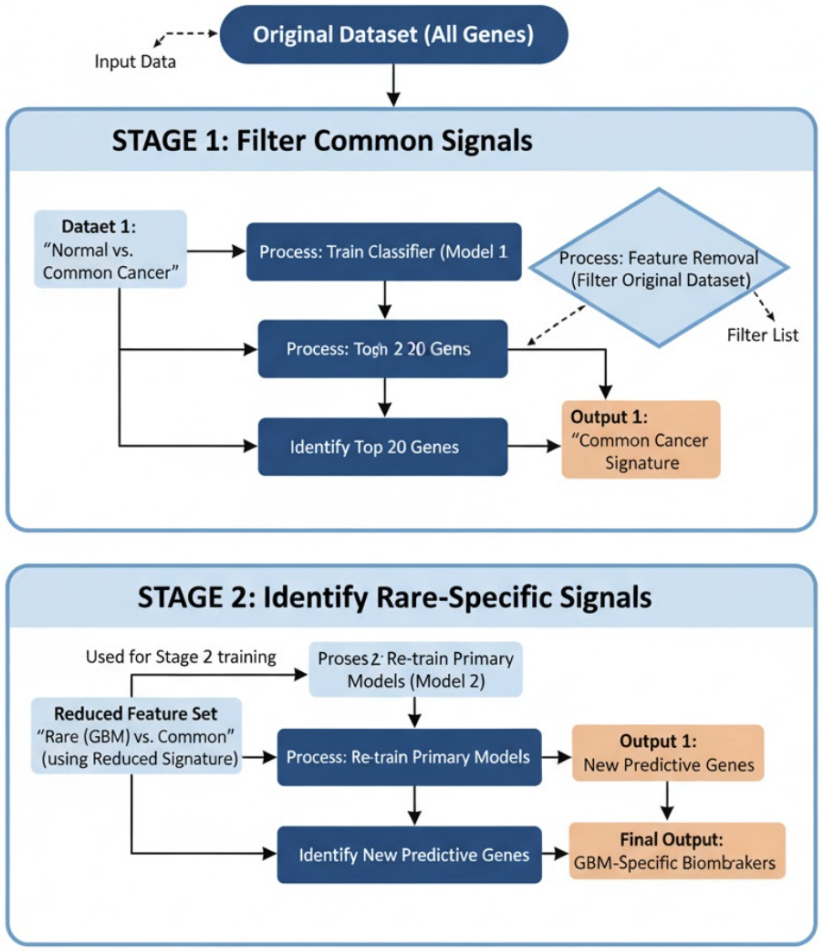

Figure B.4: Two-stage Cascade Learning framework for rare cancer signal isolation.

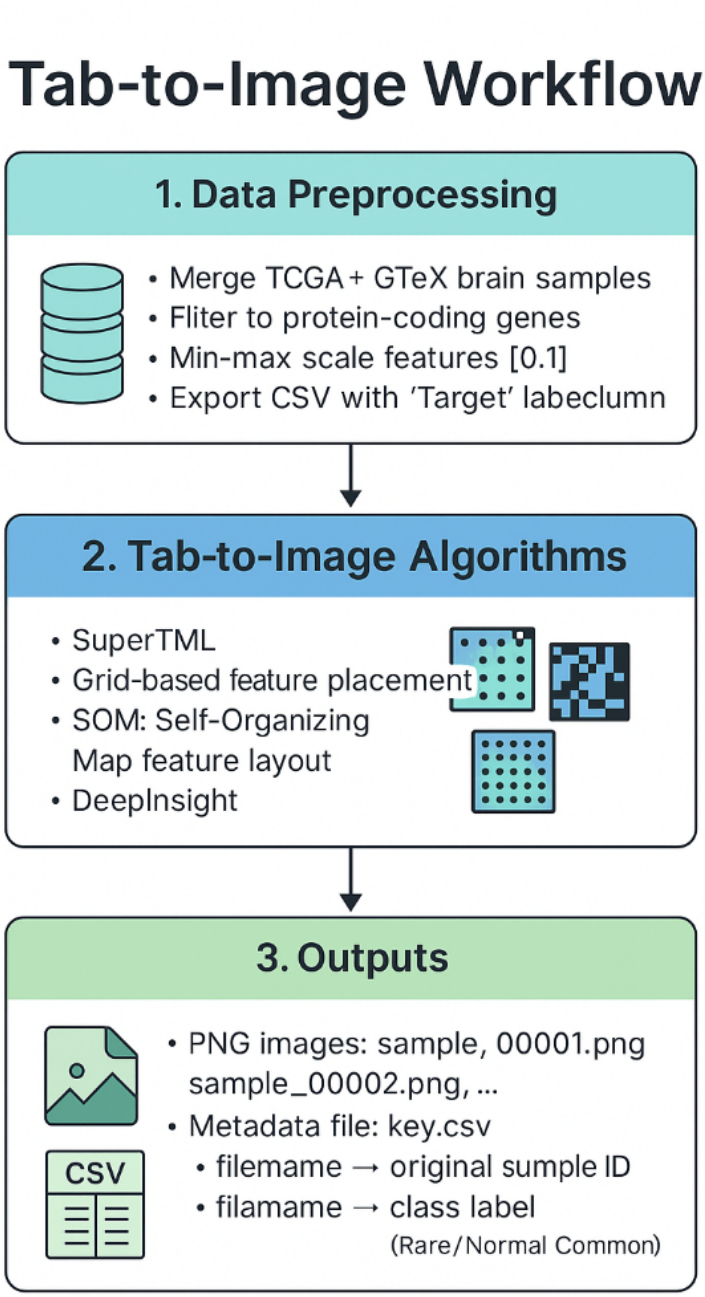

Figure B.5: Tab-to-Image conversion process for model interpretability.

Appendix C

EXTENDED PERFORMANCE RESULTS AND SHAP INTERPRETABILITY

This appendix expands performance metrics and SHAP interpretability results referenced in Chapters 4–5.

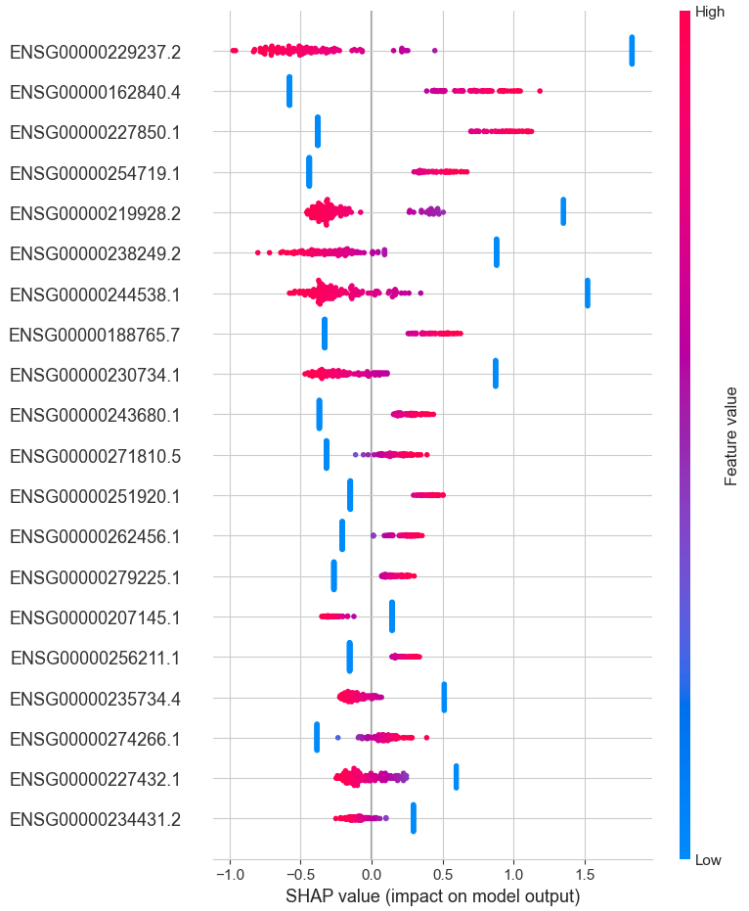

Figure C.1: SHAP summary plot for Scope 2: Rare vs Common (LGG).

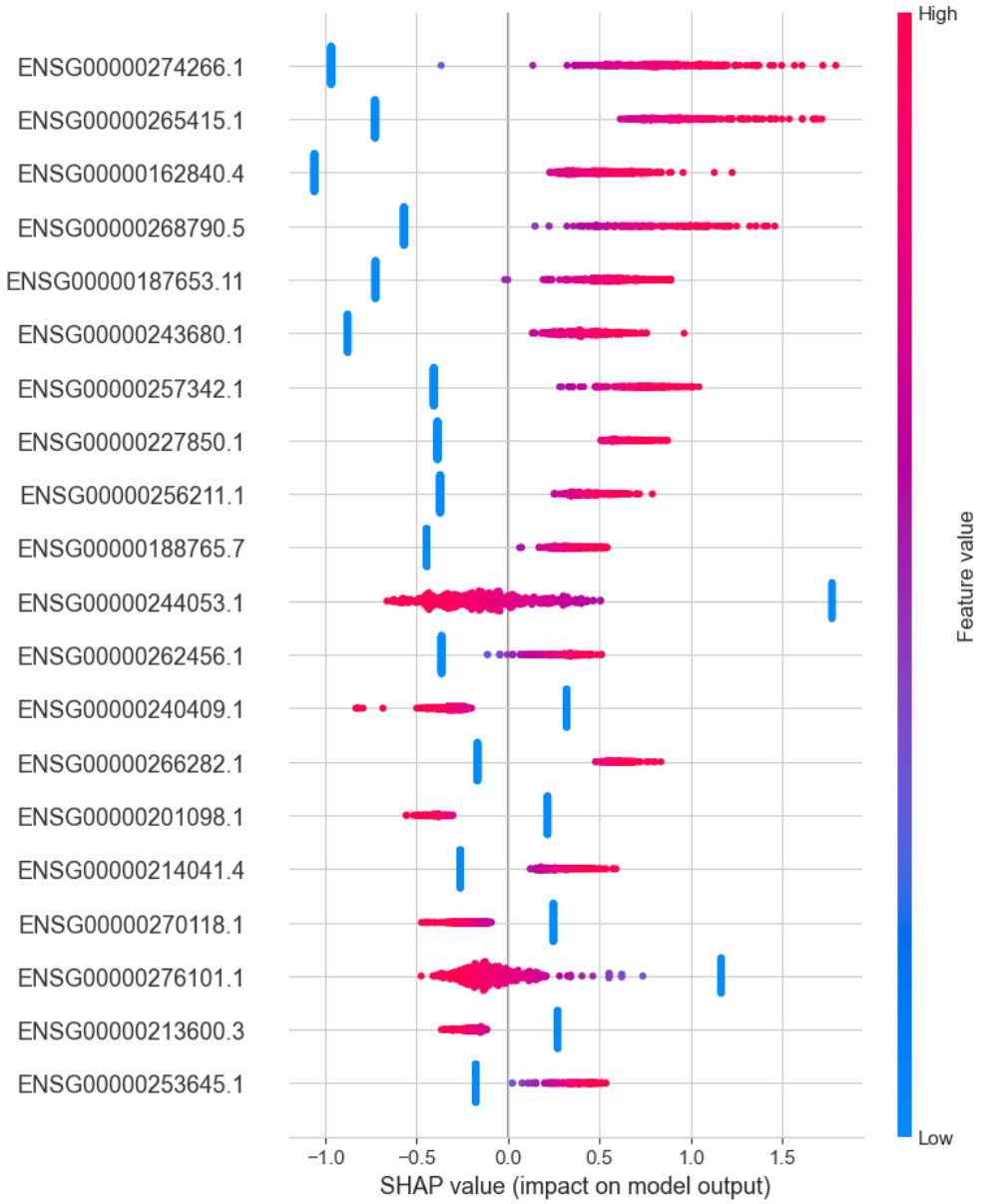

Figure C.2: SHAP summary plot for Scope 3: Rare vs Normal + Common (All Brain Tissues).

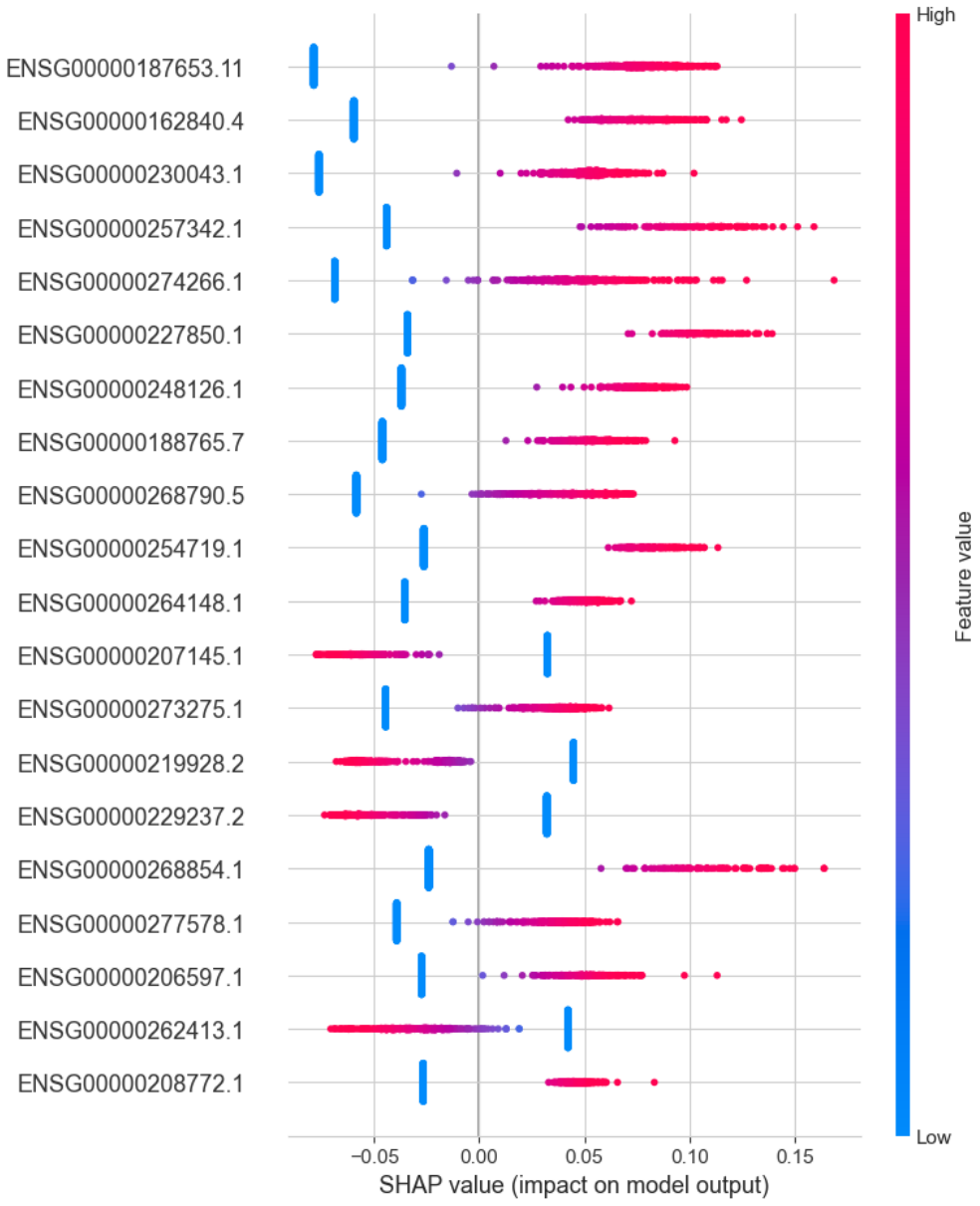

Figure C.3: SHAP summary plot for Scope 4: Rare vs All Other Gene Expression Data.

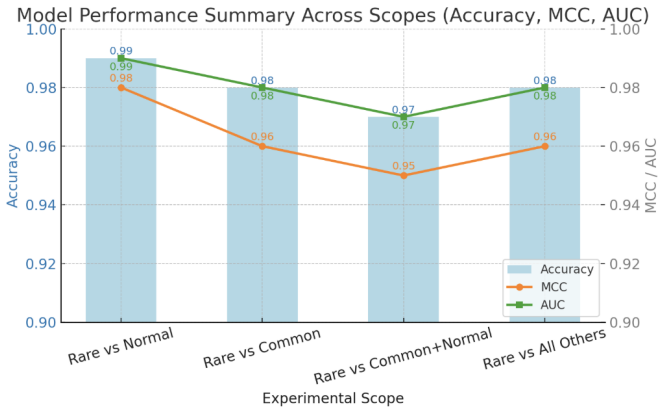

Figure C.4: Accuracy, MCC, and AUC across the four classification scopes, showing consistently strong GBM predictive performance even as task difficulty increases.

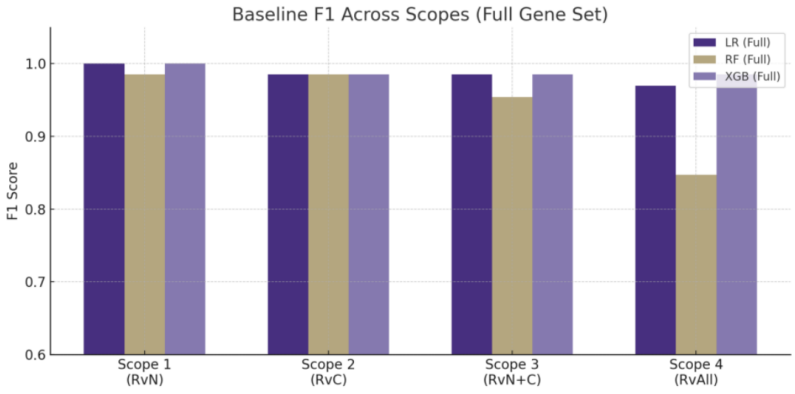

Figure C.5: Baseline F1 scores across all four experimental scopes using the full gene set.

Appendix D

EXTENDED CONSENSUS BIOMARKER SCOPE-WISE PANELS

This appendix presents scope-wise consensus biomarker rankings, model recurrence heatmaps, and weighted score distributions that support the summary provided in Chapter 5. These materials offer full transparency into the robustness analysis pipeline by showing how features were recovered across algorithms, sampling schemes, and biological settings. Interpretation of the biological relevance of these panels appears in the main text; the figures here serve as supplementary evidence of reproducibility.

To avoid repetition, detailed pathway interpretation is not repeated here. Readers are referred to Chapter 5 for biological analysis of these marker families and to Appendix E for related cascade learning results that demonstrate specificity filtering.

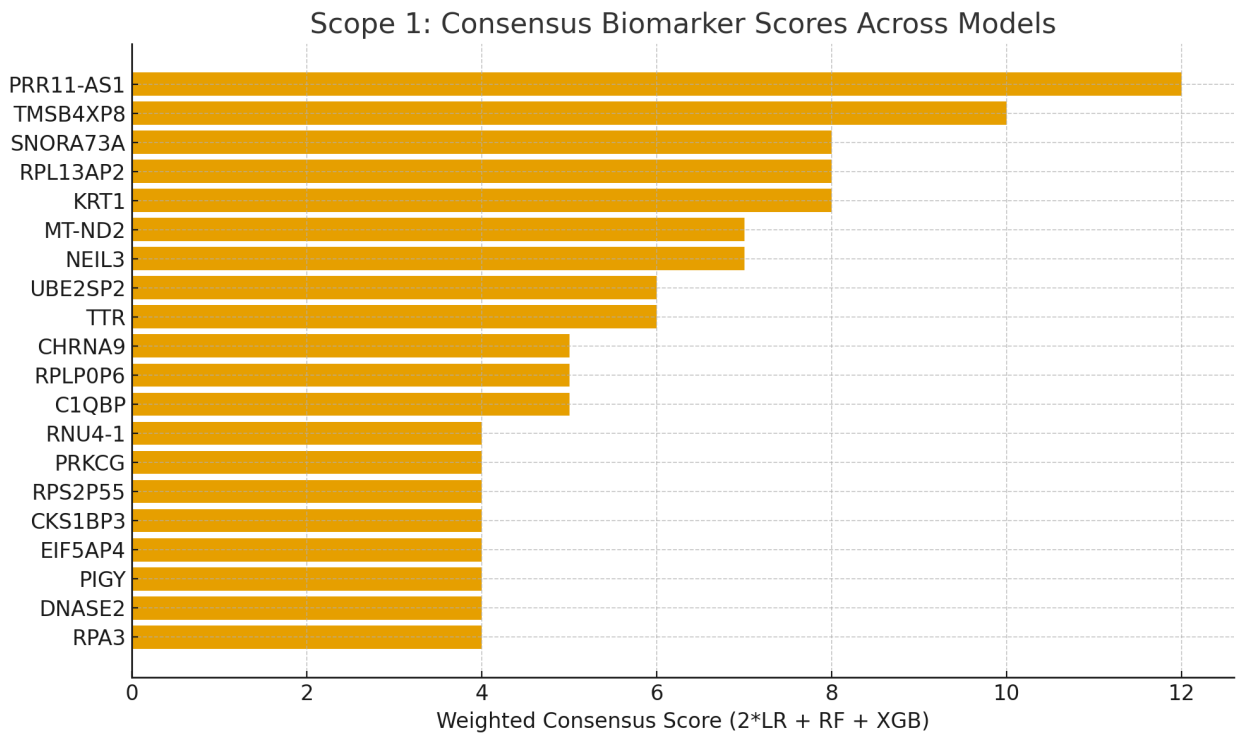

Figure D.1: Weighted consensus scores for the top twenty genes in Scope 1 (Rare vs Normal).

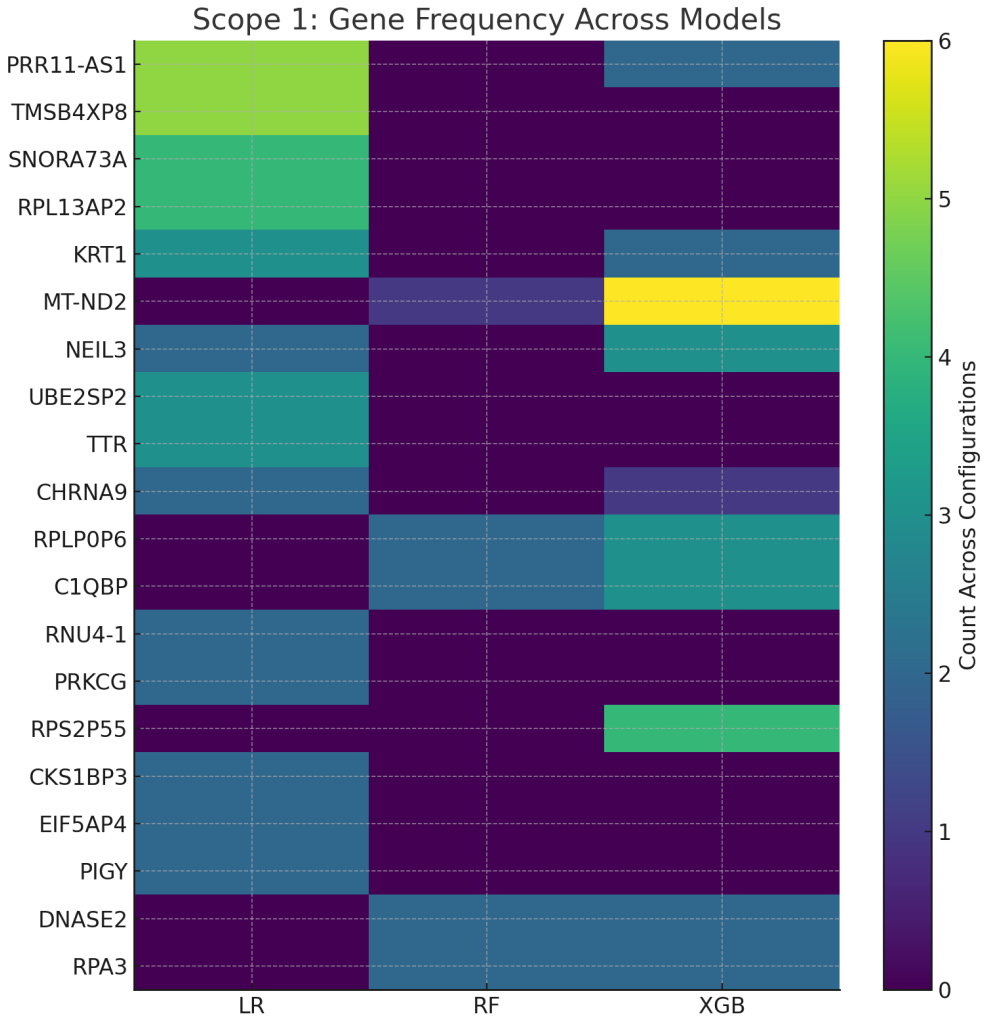

Figure D.2: Model recurrence heatmap for Scope 1, showing how often each gene is selected across LR, RF, and XGB configurations.

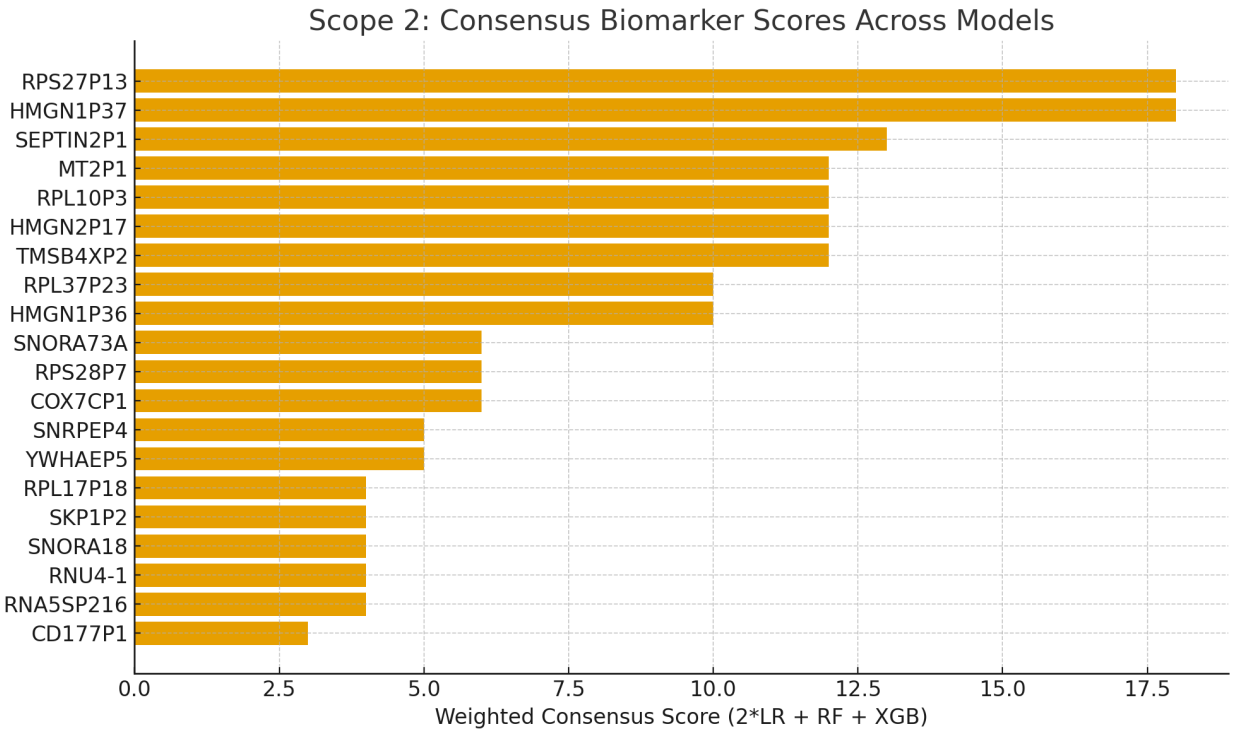

Figure D.3: Weighted consensus scores for the top twenty genes in Scope 2 (Rare vs Common).

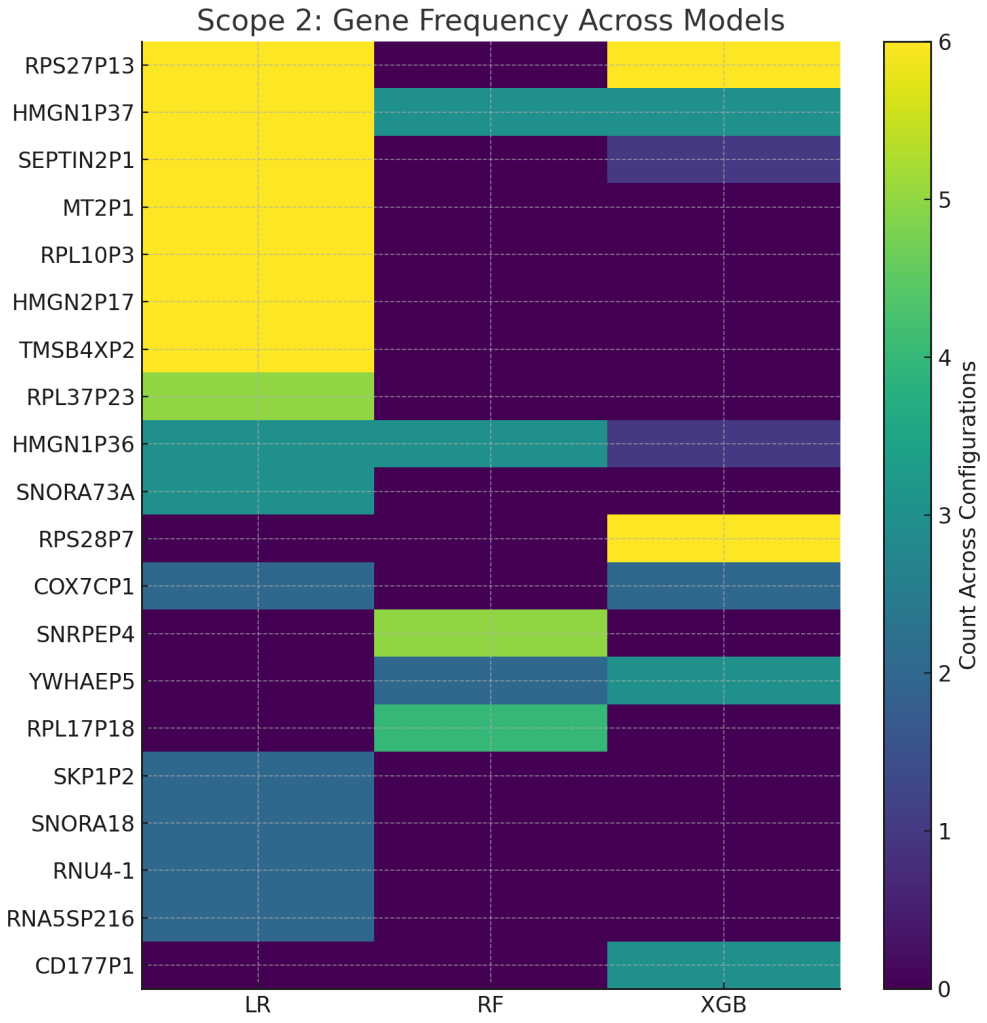

Figure D.4: Model recurrence heatmap for Scope 2 across LR, RF, and XGB, illustrating gene stability under glioma-to-glioblastoma comparison.

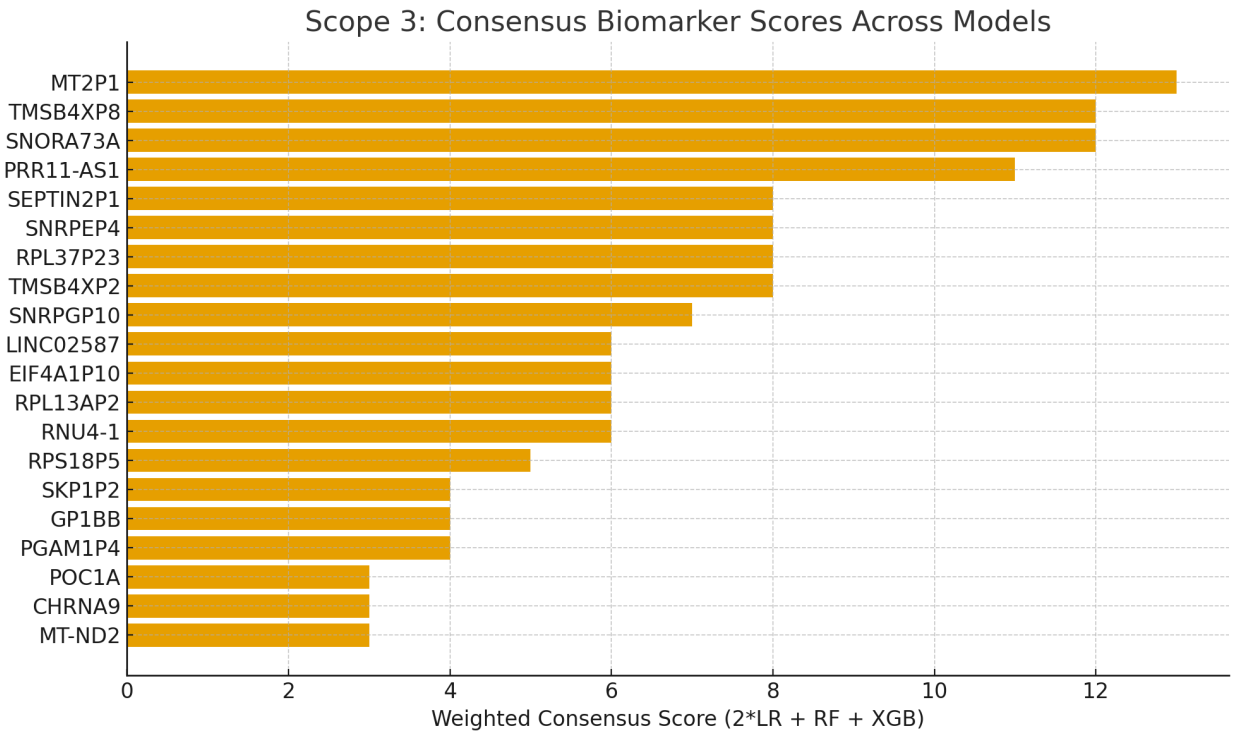

Figure D.5: Weighted consensus scores for the top twenty genes in Scope 3.

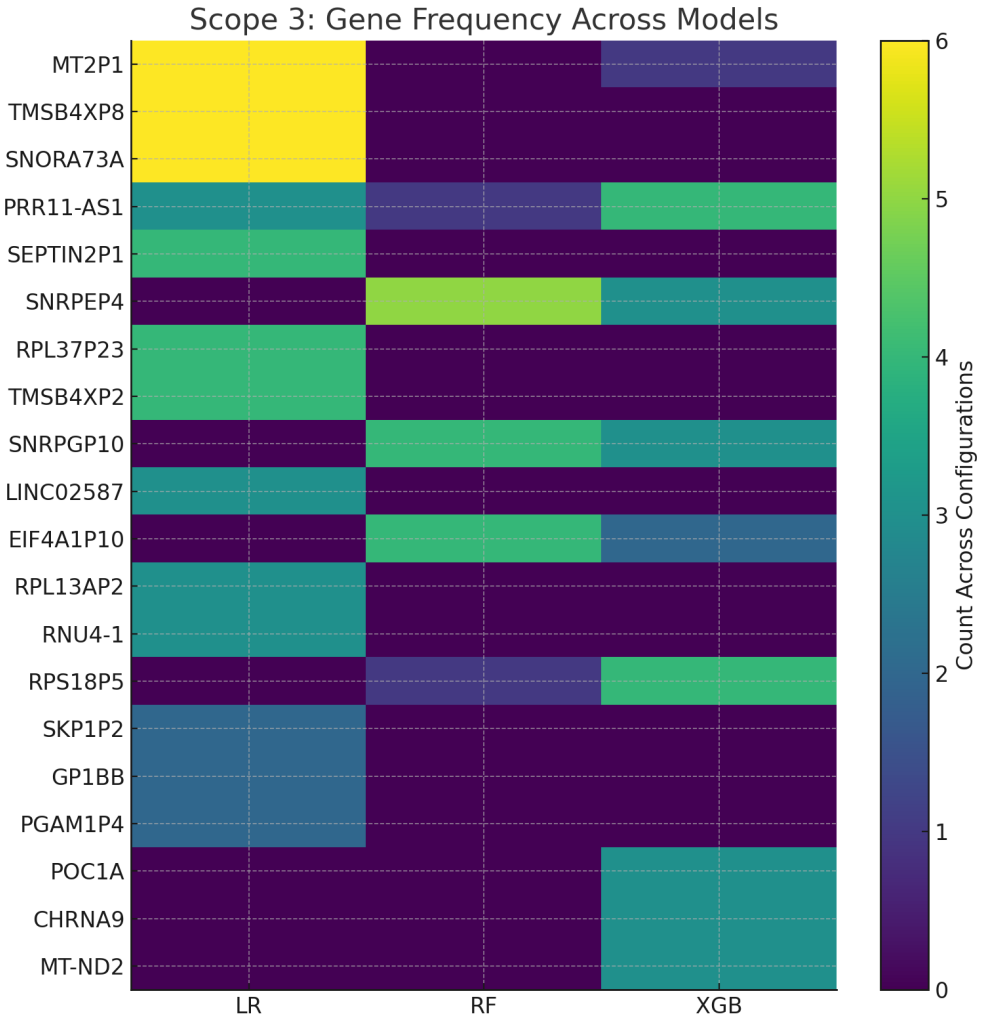

Figure D.6: Model recurrence heatmap for Scope 3, demonstrating stability of rare-cancer signal under biologically mixed negative class conditions.

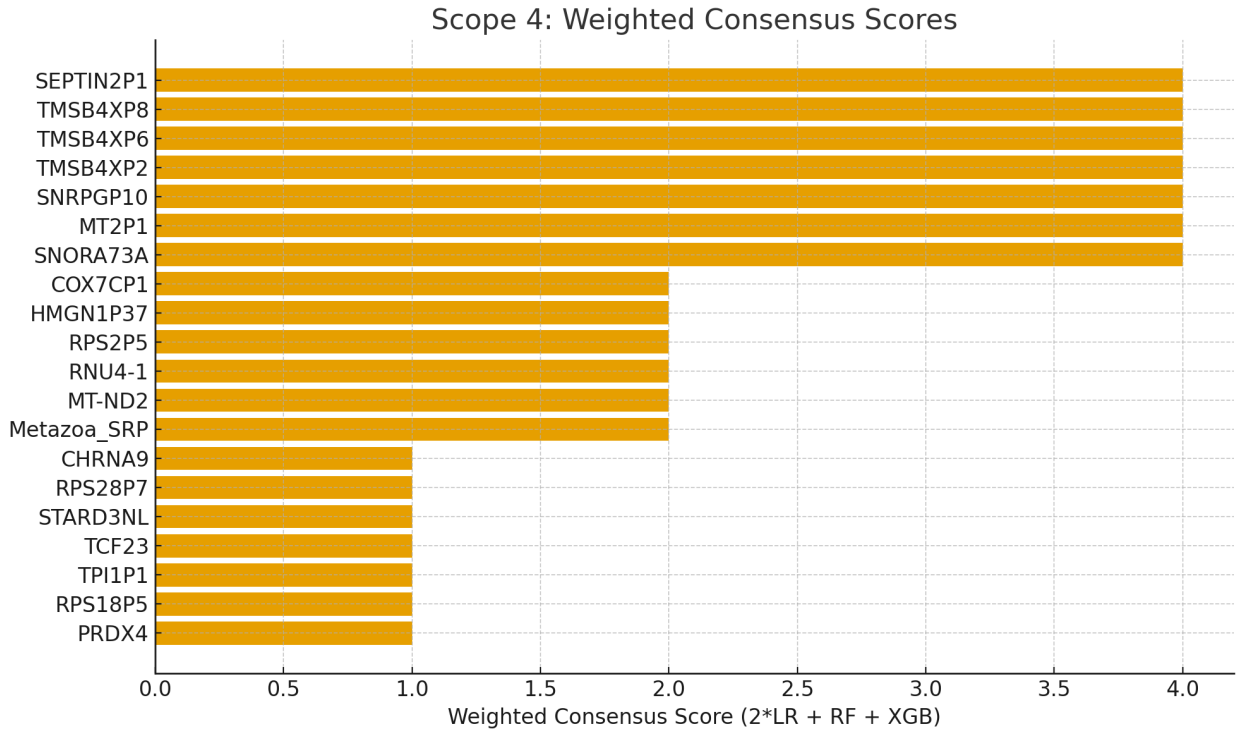

Figure D.7: Weighted consensus scores for the top twenty genes in Scope 4 (Rare vs All Other Cancers and Normal Tissues).

Figure D.8: Model recurrence heatmap for Scope 4 showing consistent biomarker recovery under pan-cancer comparison.

Appendix E

CASCADE LEARNING

This appendix reports the detailed overlap statistics and visual results for the Cascade Learning experiments described in Section 3.9.

Table E.1: Effect of cascade filtering on gene overlap and specificity.

Experiment	**Shared Before**	**Shared After**	**Rare-Specific Gain**	**F1 Change**
Rare vs Common	10	0	+10	1.00 to 0.99
Rare vs Normal	2	0	+18	1.00 to 1.00

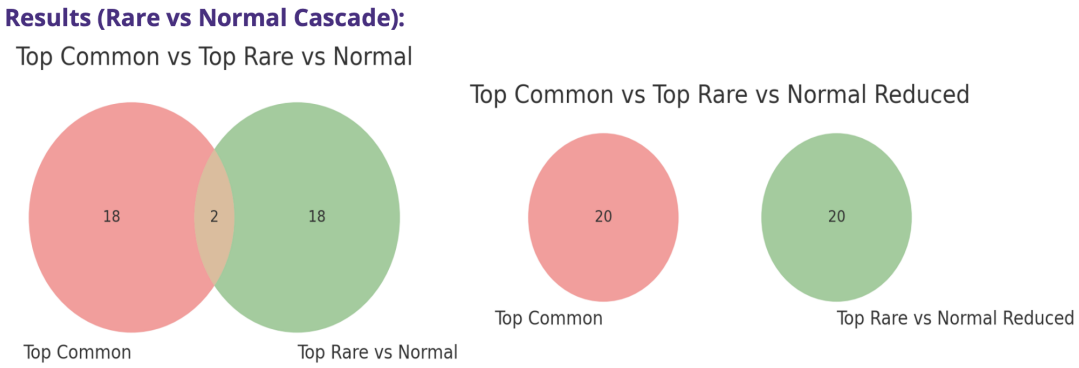

Figure E.1: Cascade results for Scope 1 (Rare vs Normal). The left Venn diagram shows overlapping top 20 genes before filtering; the right diagram shows disjoint panels after removal of shared features.

Appendix F

TAB2IMAGE

This appendix provides additional Tab2Img visualizations that complement the summary in Section 4.7. The figures show how SOM and DeepInsight layouts capture the progression from normal brain tissue to common brain cancers and rare GBM.

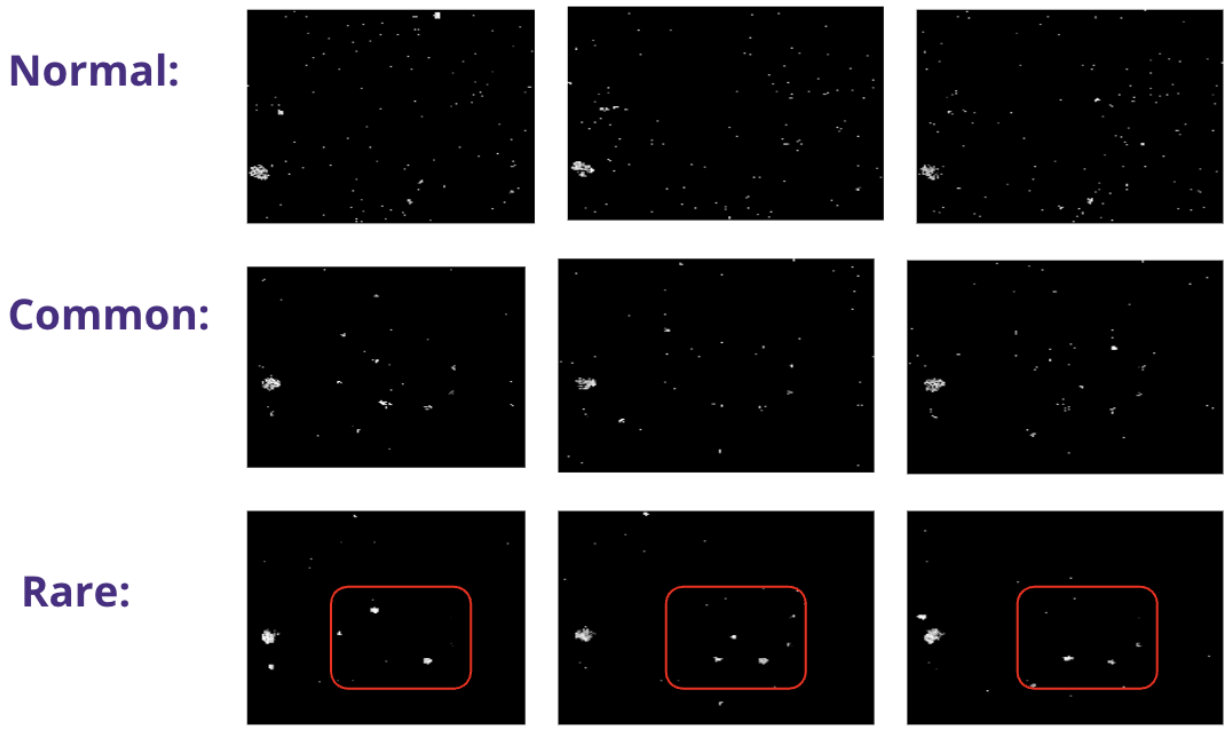

Figure F.1: Representative Tab2Img visualizations using Self-Organizing Maps (SOM), illustrating increasing activation density and cluster compactness from normal tissue to common cancers to rare cancer GBM.

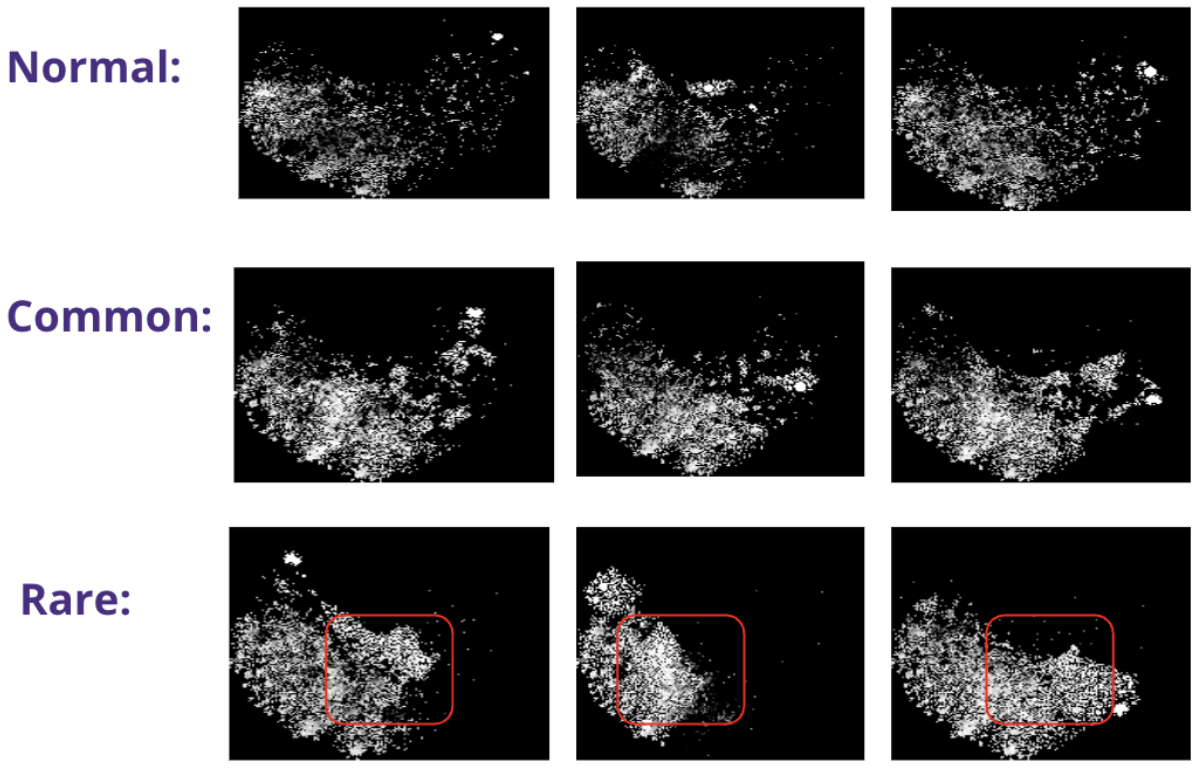

Figure F.2: Representative Tab2Img visualizations using DeepInsight, showing a progression from diffuse low-intensity patterns in normal tissue to structured intermediate activation in common cancers and dense focal activation in rare cancer GBM.