Decoding Breast Cancer Mutational Signatures: A Hybrid ElasticNet–XGBoost Approach Using Gene Expression Data

Omji Porwal[1], Kamal Upreti[2,*], Pravin R. Kshirsagar[3], Sarika Panwar[4], Anurag Sharma[5], Ganesh V. Radhakrishnan[6], Rituraj Jain[7]

[1]*Faculty of Pharmacy, Qaiwan Research Center, Qaiwan International University, Sulaymaniyah, Kurdistan, 46001, Iraq*
omji.porwal@uniq.edu.iq

[2]*Department of Computer Science, Christ University, Delhi NCR, Ghaziabad, 201002, India*
kamalupreti1989@gmail.com

[3]*Department of Electronics & Telecommunication Engineering, J D College of Engineering & Management, Nagpur, Maharashtra, 441501, India*
pravinrk88@yahoo.com

[4]*Departmemt of Electronics and Telecommunication Engineering, MIT Academy of Engineering, Pune, India*
sarika.panwar@mitaoe.ac.in

[5]*School of Electrical and Electronic Engineering, Newcastle University, NE1 7RU, Singapore*
anurag.sharma@newcastle.ac.uk

[6]*Kalinga School of Management, Kalinga Institute of Industrial Technology, Bhubaneswar, 751024, India*
vrkris2002@gmail.com

[7]*Department of Information Technology, Marwadi University, Rajkot, Gujarat, 360003, India*
jainrituraj@yahoo.com

**Corresponding Mail ID : kamalupreti1989@gmail.com*

ABSTRACT

TP53, PIK3CA, and MUC16 are somatic mutations that are useful in breast cancer progression and prognosis, but direct mutation profiling based on sequencing is not always practicable in practice. The data about gene expression can contain indirect transcriptomic patterns linked with mutational underlying states. This paper proposes an expression-based machine learning model to predict the status of mutations using METABRIC breast cancer cohort. Instead of directly estimating genetic changes, the suggested method estimates statistical relationships between transcriptomic phenotypes and binary somatic mutation states. A multi-stage gene features selection pipeline using variance filtering, mutual information ranking, and correlation pruning was used to reduce the number of genes (19,000). A hybrid predictive architecture was trained using these features that combined ElasticNet logistic regression and XGBoost that allowed balancing between linear regularization and nonlinear interaction modeling. The hybrid model with a combination of five-fold stratified cross-validation yielded mean ROC-AUC of 0.94 (TP53), 0.92 (PIK3CA), and 0.90 (MUC16) with the stability of the calibration and equal error rates. Coefficient analysis and SHAP-based explanations were used to investigate the interpretability of the models to describe the expression patterns on mutation status. The suggested framework is a hypothesis-generating, complementary method of transcriptomic analysis, which must be reevaluated by external validation to determine the wider generalizability.

https://doi.org/10.36001/ijphm.2026.v17i1.4714

1. INTRODUCTION

Breast cancer is a heterogeneous disease at the molecular level that has a complicated pathogenesis of genetic

mutations and down-stream transcriptional deregulation. The somatic mutations of major regulatory genes have an impact on cellular growth, proliferation and survival, which ultimately affect the tumor behavior and clinical outcomes. Gene expression analysis has traditionally been an effective mechanism of defining these downstream effects because it captures transcriptional fingerprints, correlated to disease status, prognosis and therapeutic response (Brady et al., 2022). With the advent of microarray technology in the mid-1990s, the new high-throughput transcriptomic profiling has made it possible to measure thousands of genes simultaneously, which has resulted in the first expression-based diagnostic and prognostic signatures, such as the first leukemia expression classifier to be proposed in 1999. These signatures have since been the core of cancer risk stratification especially when conventional clinicopathological signatures are not sufficient to be used alone.

In the last twenty years, many prognostic gene-expression signatures have been suggested in various types of cancers (Qian et al., 2021). In breast cancer, nodal involvement, proliferation markers, and the status of hormone receptors, in the combination of surgery with adjuvant endocrine or chemotherapeutic treatment, usually influence the treatment choices of the patients having hormone receptors and are without hormone receptors, HR-negative and with no HER2 on the tumour. Nevertheless, especially in HR+/HER2- luminal B-like tumors, these traditional predictors do not fully result in the underlying biological complexity and intratumoral heterogeneity which modulate disease progression and response to therapy (Munkacsy et al., 2022). This shortcoming has promoted the further evolution of transcriptomic protocols that provide more detailed stratification of the molecules.

TP53, PIK3CA and MUC16 are somatic mutations which are applicable in the development and prognosis of breast cancer, however, direct mutation profiling via sequencing is not always feasible in the regular workflow (Betz et al., 2025). Indirect transcriptomic trends can be provided by the gene expression information that is associated with mutational conditions. In this paper, analysis is an expression-based machine learning model to predict mutation status on the METABRIC breast cancer cohort. The proposed approach does not make direct inferences of genetic changes rather; it models statistical correlations between transcriptomic variables and binary somatic mutation labels.

Multi-stage feature selection pipeline feature filtering by variance, ranking mutual information, and pruning correlation were used to narrow down the number of genes to compact and mutation-specific features. These characteristics were trained in a hybrid predictor architecture that combined ElasticNet logistic regression with XGBoost to provide a trade-off between linear regularization and nonlinear interaction modeling.

The hybrid model performed at 0.94 mean ROC-AUC of TP53, PIK3CA (0.92 and 0.90) and a stable calibration, and balanced error rates. Coefficient analysis and SHAP-based explanations were used to investigate model interpretability to describe the pattern of expression by mutation status. The proposed framework is presented as a hypothesis-generating, complementary method of transcriptomic analysis, and external validation is required to determine how widely the findings can be generalized.

Massive transcriptomic research has also demonstrated the potential and limitations of expression-based classification of breast cancer. Horr and Buchler (2021) discovered seven Breast Cancer Consensus Subtypes (BCCS) using 5,950 samples and showed that whole-transcriptome analysis may refine the taxonomy of a disease. However, they also highlighted significant limitations, such as the requirement of analytically validated assays that could be used on formalin-fixed and paraffin-embedded (FFPE) tissue, the requirement of accurate determination of estrogen receptor (ER) status, and the possibility of assigning ambiguous subtypes due to intratumoral heterogeneity. These are the challenges that point to the complexity of applying complex transcriptomic signatures to clinical practice, in spite of their biological relevance.

Simultaneously, artificial intelligence (AI) and machine learning solutions have become an object of growing interest in the research of breast cancer, not just in the imaging context, but also in the context of molecular data analysis. The recent literature has investigated applying machine learning to determine mutation-associated or prognosis-related patterns in genomic and transcriptomic data. Odhiambo et al. (2023) concentrated on mutational signatures and their relationship with genetic profiles in breast cancer, whereas the rest of the work has combined clinical variables with computational software, including Geneshot, Phenolyzer, and Cytoscape, to analyze the relationship between genes and diseases. Oncogenic mutations, such as BRCA1, BRCA2, and TP53, have also been well-established, and it has been established that these mutations have demonstrable downstream consequences in terms of gene expression programs, prompting interest in computational models that can be used to relate transcriptomic patterns to mutation status.

Other machine learning investigations have shown that it is possible to come up with clinically relevant gene-expression signatures. Thalor et al. (2022) implemented recursive elimination of features and XGBoost to obtain transcriptomic signatures to differentiate between triple-negative breast cancer (TNBC) and Kim et al. (2025) used next-generation sequencing-based expression data and several machine learning models to determine relapse-related gene signatures in TNBC. They reported high predictive performance of their optimized models and sensitivity of 0.8750, specificity of 0.9231 and area under ROC curve estimated as 0.9087, which

is backed by Kaplan-Meier survival analysis. These papers emphasize the usefulness of machine learning in the identification of prognostically informative patterns of expression, and demonstrate as well that the vast majority of current methods are interested in either subtype classification or outcome prediction but not direct modeling of mutation status.

Although the state of the art in genomic profiling methods has been achieved, the existing breast cancer prognostic assays are still, to a great extent, relying on clinicopathological aspects and preset expression panels, which might not be adequate in capturing the dynamic nature of the relationship between somatic mutations and downstream transcriptional activity. Secondly, although gene expression represents the functional outcomes of genetic change, this is not a direct measure of mutations, only the functional outcome. This difference is one that should not be ignored because only statistical relationships can be discovered using expression-based models, but not causal genetic pathways. In this respect, it will be desirable to have well-constructed computational strategies that utilize transcriptomic data to investigate mutation-related expression patterns without excessive biological or clinical speculation.

Clinically, various kinds of prediction errors have dissimilar implications. In mutation-status prediction, false-negative error, in which mutation-positive tumors are falsely predicted to be non-mutated, can be more damaging than false positives as it can constrain the downstream molecular stratification or therapeutic consideration. Accordingly, this study emphasizes evaluation metrics that capture class-specific performance and calibration, rather than relying solely on overall accuracy. These considerations informed both model selection and the use of complementary performance measures to reflect clinically relevant trade-offs.

Our hybrid machine learning proposal in this study will be a predictive model of the somatic mutation status (TP53, PIK3CA, and MUC16), based on gene expression profiles of the METABRIC cohort. The models predict statistical correlations between transcriptomic characteristics and binary mutation labels, as opposed to predicting genetic changes directly. The framework combines ElasticNet logistic regression with XGBoost, which provides a balance between linear regularization and nonlinear interaction modelling, and supports model-level interpretability with SHAP-based analysis that appears in the context of hypothesis generation. The main aim of the research is to determine whether it is possible to learn compact, mutation-associated and expression signatures with strong reproducibility on a large-scale transcriptomic data, thus offers a complementary computational view into the heterogeneity of breast cancer on the molecular level. A key limitation of the current research is the lack of external validation through independent sets. Despite the impressive predictive accuracy provided by internal cross-validation, it is crucial to verify the findings using external datasets. Future work should also be in external validation and clinical translation. The clinical relevance of the suggested model must be cautiously assessed if there is no external validation.

While sequencing-based detection is the current gold standard for mutation analysis, prediction based on gene expression can be considered an alternative approach in situations where sequencing facilities are not available. The computational approaches discussed above may be useful for initial screening and testing of hypotheses.

2. MATERIALS AND METHODS

This study meant to model the statistical associations between patterns of gene expression and somatic mutation status, not to make an inference on genetic alterations. Somatic mutations are fixed DNA phenomena, and gene expression is the downstream transcriptional action which depends on genetic and epigenetic factors. To this end, it is proposed that mutation status serves as a binary target variable, and expression profiles as predictive features, which capture expression signatures, which are related to the presence or absence of particular mutations. There is no causal implication of gene expression and mutation occurrence. The supervised binary classification is established as the prediction task, and it is also independent of the target mutation (TP53, PIK3CA, and MUC16) with the aim of estimating the likelihood of mutation presence based on a transcriptomic profile.

2.1. Data Acquisition and Description

The present work uses the data of Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) which is a popular and high-quality public database that offers the matched transcriptomic and somatic mutation annotations on the cases of breast cancer (Breast Cancer Gene Expression Profiles (METABRIC), n.d.). As an open source dataset, it is commonly used in computational oncology studies because its preprocessing is standardized, extensive clinical annotation exists, and it helps to achieve transparency and reproducibility in data-based research (Mukherjee et al., 2018). The METABRIC cohort consists of 1,980 primary tumors of breast with measured normalized genes expression of around 19,000 protein-coding genes. Besides transcriptomic data, annotated somatic mutation data is available on some of the most important oncogenic drivers, such as TP53, PIK3CA, MUC16, BRCA1, BRCA2, GATA3, or MAP3K1. Notably, mutation data are not taken as input features in this work, but solely as binary target labels namely whether a particular somatic mutation is present or not. Three mutation targets—*TP53*, *PIK3CA*, and *MUC16*—were selected due to their well-established involvement in breast cancer biology, including DNA damage response, PI3K/AKT signaling, and cell adhesion or metastatic regulation,

respectively. For each target, mutation status was encoded as a binary variable $y_i \in 0,1$, while the corresponding gene expression profile was represented as a high-dimensional feature vector Xi where $[x_{i1}, x_{i2}, \dots, x_{in}]$, where n≈19,000. This formulation explicitly models mutation-status prediction as a supervised classification task based on transcriptomic features, without implying direct inference of genetic alterations.

The proportion of mutations is realistic as the proportion of TP53 in the cohort is 34.2%, PIK3CA is 27.5%, and MUC16 is 13.8%, which are comparable to the reported mutation levels in the population with breast cancer. Data partitioning was done in a stratified 80:20 train-test to guarantee representative sample sizes and decrease sampling bias, and to guarantee equal distribution of mutation among subsets. Metadata were checked in an organized manner and they have been verified to exclude duplicated samples, batch artifacts and non-numeric anomaly. The identifiers of genes were assigned based on HUGO Gene Nomenclature Committee (HGNC) conventions to be compatible with downstream pathway databases, including KEGG and Reactome (Gale et al., 2021). The resultant data format, which is highly dimensional in comparison to sample size (n ≫ m), is known to have significant issues to do with overfitting, redundancy of features and interpretability. This feature renders the METABRIC dataset specifically appropriate in assessing hybrid learning models that explicitly trade dimensionality reduction, predictive accuracy and model interpretability, which are the key methodological goals of the present study.

2.2. Data Preprocessing

To achieve the reliability of the analytical procedure and reduce bias in high-dimensional gene-expression research, proper preprocessing is required. Since METABRIC dataset is heterogeneous and has large scale, there are several systematic operations carried out to clean, normalize, and standardize the raw features prior to modeling. The general preprocessing process featured four major steps, namely, data cleaning, normalization, outlier control, and rebalancing of the classes. The main goal was to maintain biological variability and minimize the statistical noise and be able to compare across samples.

2.2.1. Missing Value Imputation

Despite the fact that the METABRIC data is mostly curated, initial verification revealed that there is a small percentage (less than 0.5%) of entries that are missing in various gene-expression features. To overcome this, a feature-wise median imputation was used which is more resistant to skewed distribution compared to mean substitutes.

Formally, for each gene j and sample i, the imputation function is defined as:

$$x_{ij}^{*} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed} \\ \tilde{x}_j = \text{median}(x_j), & \text{if } x_{ij} \text{ is missing} \end{cases}$$

where x_{ij}represents the observed value of gene $j$for patient i, and $\tilde{x}_j$denotes the median expression value of gene $j$across the entire dataset. This operation ensured numerical completeness while maintaining original distributional properties.

2.2.2. Data Normalization

The raw gene-expression values being of different orders of magnitude, z-score normalization was used to bring the features to a normalized range. This transformation removes bias in training models as genes with larger raw magnitudes will not dominate gradient updates and regularization terms.

The normalization process can be mathematically represented as:

$$z_{ij} = \frac{x_{ij}^{*} - \mu_j}{\sigma_j}$$

where $\mu_j$and σ_j are the mean and standard deviation of gene j respectively calculated over all samples. The outcome is a normalized feature space with a mean of zero and unit variance of every gene. It is also conducive to biological comparability because the procedure enables changes occurring at an expression level to be described in relative, but not absolute terms.

2.2.3. Outlier Detection and Removal

In order to protect against possible sequencing or measurement artifacts, an outlier analysis was performed based on Interquartile Range (IQR) filtering on every gene (Khan et al., 2019). The extreme observations were considered to be values that are more than 1.5xIQR above the third quartile or less than the first quartile. Formally:

$$\text{Outlier}(x_{ij}) = \begin{cases} 1, & \text{if } x_{ij} < Q1 - 1.5(Q3 - Q1) \text{ or } x_{ij} > Q3 + 1.5(Q3 - Q1) \\ 0, & \text{otherwise} \end{cases}$$

Around 2.3% of total data points were identified as the potential outliers. Given that the dynamic range of biological expression profiles inherently has large values, however, such values were not simply eliminated but rather winsorized to the corresponding quantile ends. This plan avoids the over truncation of biologically valid but extreme responses of genes.

2.2.4. Class Imbalance Management

The nature of mutation prediction problems is characterized by class imbalance where mutated samples are in the minority. In the case of the METABRIC data, MUC16 mutations were found in only 13.8% of the patients, which would create the risk of biased learning to the majority (non-mutated) group. To fill this, the Synthetic Minority Oversampling Technique (SMOTE) was used to come up with synthetic positive samples using nearest-neighbor-based

interpolation on the feature space of the minority class (Shi et al., 2023). For a given sample x_i and one of its k-nearest neighbors x_{nn}, a synthetic sample x_{new} is generated as:

$$x_{new} = x_i + \delta \times (x_{nn} - x_i)$$

where $\delta \sim U(0,1)$.

The method generates synthetic but realistic points on the feature-space manifold thereby enhancing balance of classes without altering original variance structure. After balancing, mutation ratios were brought to about 50:50 of all three target genes, which greatly enhanced sensitivity and recall in next-generation classification. In order to avoid information leak, synthetic samples were only generated inside training folds when performing the cross-validation and were not added to the validation or test partitions.

2.2.5. Data Partitioning and Reproducibility Control

All of the preprocessing procedures were run under constant random seeds (42) to guarantee reproducibility and recorded in the Jupyter notebook environment. Data partitioning was based on a stratified 80:20 rule, such that there was an equal representation of mutation-containing and mutation-absence samples in each of the subsets. A similar preprocessing pipeline that was executed in scikit-learn (v1.4.2) captured each modeling phase, ensuring consistent results when repeated in multiple experimental runs. Table 1 presents a summary of pre-processing and post-processing data statistics, such as the number of samples, the percentage of missing values, and the measures of spread.

Stage	Samples	Features	Missing Values (%)	Variance Retained	Mutation Ratio (Positive %)
Raw dataset	1980	19,000	0.47	100 %	TP53: 34.2%, PIK3CA: 27.5%, MUC16: 13.8%
Post-imputation	1980	19,000	0.00	99.8 %	same
Post-normalization	1980	19,000	0.00	99.6 %	same
Post-SMOTE balancing	2640 (synthetic added)	19,000	0.00	99.2 %	TP53: 50, PIK3CA: 50, MUC16: 50

Table 1. Data quality and transformation summary

The preprocessing pipeline ensured the integrity, comparability, and balance of the input features. The statistical normalization, outlier winsorization, and synthetic balancing allowed the data to achieve the two objectives of maintaining relative transcriptomic structure and having analytical strength and cross-sample comparability.

2.3. Feature Selection and Dimensionality Reduction

The high-dimensional transcriptomic data are usually characterized by a vast difference between the quantity of measured features and the quantity of usable samples, which creates high risks of overfitting and model instability. The METABRIC cohort has around 19,000 gene expression variables per 1,980 samples (n>>m), which translate to an underdetermined learning problem which requires explicit dimensionality reduction. In this regard, feature selection was used to enhance the numerical stability, minimize redundancy and increase the interpretability of the model. Notably, the feature-selection process is not intended to determine biological causality but only to detect statistically informative expression features by mutation status. The combination of three-stage hybrid filtering strategy (filtering based on variance, filtering based on information theory and relevancy, and filtering based on correlation) was adopted. This gradual mode allows the systematic reduction of dimensionality whilst preserving the expression characteristics that bring about substantial value to predicting mutation-status.

2.3.1. Stage I – Variance Thresholding

Genes that show little variation of expression in samples do not add much to the discrimination of a given class. Variance thresholding was used to eliminate such features that are noise-dominant (Dinalankara & Bravo, 2015). The variance of each gene $X_j$was computed as:

$$\mathrm{Var}(X_j) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2$$

Genes with variance below a threshold $\tau_v = 0.01$, selected empirically based on the variance distribution, were removed. This step eliminated approximately 35% of near-constant genes, reducing the feature set to ~12,200 genes while preserving over 98% of overall expression variability. The objective of this stage was numerical efficiency rather than biological interpretation.

2.3.2. Stage II – Mutual Information (MI) Ranking

After the filtering of the variance, mutual information (MI) was applied to measure the dependence between every gene expression variable and its associated binary mutation label (Mallik & Zhao, 2017), (Posta & Győrffy, 2025). MI unlike linear correlation, reveals both linear and nonlinear relationships. For each gene $X_j$and mutation indicator Y:

$$I(X_j;Y) = \sum_{x_j \in X_j}\sum_{y \in Y} p(x_j,y)\log\frac{p(x_j,y)}{p(x_j)p(y)}$$

Genes were ranked by MI score, and the top 5% (approximately 600–700 genes per mutation target) were retained using the *mutual_info_classif* function in scikit-learn. This step identifies expression features with the strongest **predictive association** to mutation status, without implying mechanistic or causal relevance. High-MI genes such as *BRCA1*, *PTEN*, *FOXA1*, *GATA3*, *AKT1*, *STAT3*, and *ESR1* were consistently observed across mutation targets, reflecting their known involvement in transcriptional regulation and oncogenic signaling.

2.3.3. Stage III – Correlation Pruning

The presence of high inter-features correlation may lead to instability of the regression coefficients and inflated model variance. Pearson correlation coefficients between each combination of the selected genes were calculated to overcome the problem of multicollinearity:

$$r_{ij} = \frac{\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{N}(x_{kj} - \bar{x}_j)^2}}$$

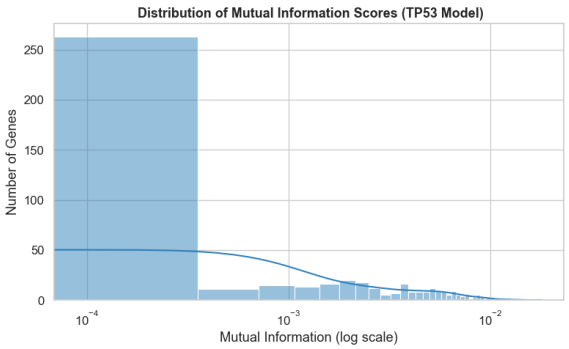
Gene pairs with $| r_{ij} |$ was at least 0.85 were deemed redundant and the feature with MI lower was eliminated. This operation minimized multicollinearity by nearly 60% and provided a small but informative set of about 40 genes per model of mutation. The resulting correlation heatmap shows that there is little redundancy of residual values of retained features, which indicates the statistical orthogonality of the final feature space.

2.3.4. Functional Contextualization of Selected Features

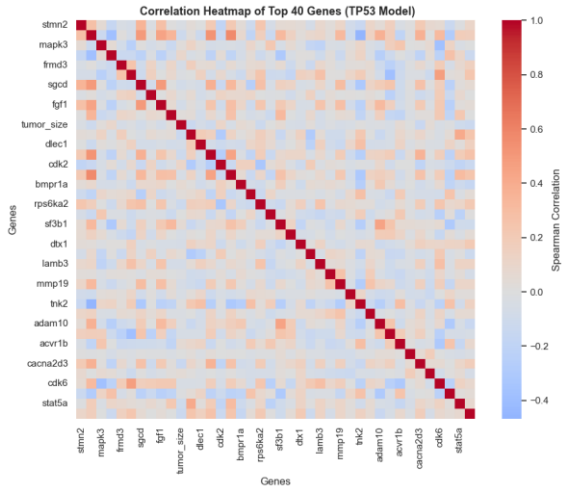
To put the statistically selected genes into context, the retained sets of features were cross-referenced with curated KEGG and Reactome pathway annotations. The resulting overlap (78% TP53, 74% PIK3CA, 69% MUC16) reflects the agreement of statistically informative features and those genes that have been previously known to play a role in cancer-related pathways. Such an overlap is claimed to facilitate interpretive plausibility, but not to declare biological causation. Figure-1 depicts the process of feature-selection: (a) a long-tailed distribution of MI scores in 19,000 genes, (b) a correlation heatmap of the remaining selected features with minimal redundancy, and (c) Venn diagrams of MI-selected genes versus curated pathway gene sets.

2.3.5. Final Dimensionality and Feature Overview

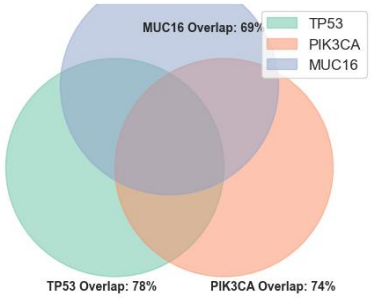
The last feature subsets were directly fed into the hybrid modeling framework. In total, the three-stage pipeline reduced dimensionality (from ~19,000 to ~40 features per mutation target) by around 99.8% which significantly reduced the computational cost with predictive signal preserved. The representative genes that were retained following selection are summarized in Table 2, with their known functional functions.



(a)



(b)



(c)

Figure 1. Feature selection and correlation refinement.

Mutation	Gene Symbol	Biological Function	Supporting Evidence
TP53	BRCA1	DNA repair, cell-cycle checkpoint	(Fu et al., 2022)
TP53	MDM2	Negative regulator of p53	(Koo et al., 2022)

PIK3CA	PTEN	PI3K/AKT pathway suppression	(Lee et al., 2019)
PIK3CA	FOXA1	Estrogen receptor co-regulator	(Seachrist et al., 2021)
MUC16	CD44	Cell adhesion and metastasis	(Senbanjo & Chellaiah, 2017)
MUC16	STAT3	Cytokine signaling, EMT regulation	(Zhang et al., 2024)

Table 2. Representative top-ranked features after hybrid selection

The suggested feature-selection plan will provide a balance between statistical rigor, dimensionality reduction, and interpretability. The resulting feature subsets introduce a stable and interpretable basis of mutation-status prediction based on the hybrid ElasticNet-XGBoost framework by explicitly considering selected genes as associative predictors, as opposed to the causal biomarkers.

2.4. Hybrid Model Construction

The designed modeling scheme is a combination of two complementary supervised learning paradigms, namely, the ElasticNet Logistic Regression (ELN) framework that can be applied to learn interpretable linear prediction (Algamal and Lee, 2015), and the Extreme Gradient Boosting (XGBoost) framework that can be used to learn nonlinear representation (Li et al., 2022). The given hybridization based on two models allows transparency and flexibility - the two qualities that are usually opposite to each other in traditional machine learning systems. The hybrid architecture (Figure 2) is made up of two independent training pipelines, each trained a different inductive bias, and a probability fusion layer, which combines predictions using a weighted ensemble algorithm. The setup is designed in such a way that global linear patterns these are represented by ElasticNet coexist with the hierarchical nonlinear dependencies represented by XGBoost, which leads to balanced generalization and interpretability. ElasticNet and XGBoost were selected to balance interpretability and nonlinear modeling capacity, rather than to exhaustively compare all possible classifiers.

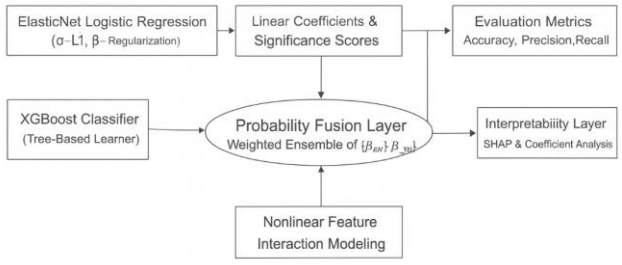


Figure 2. Schematic architecture of the hybrid modeling framework.

Hybrid modeling structure diagram of interpretable ElasticNet Logistic Regression and adaptive XGBoost classifier. In Panel (a), parallel learning streams are trained to learn different inductive biases: on the left stream, a globally linear formation is trained by L_1–L_2 regularization; and on the right stream, a nonlinear formation between features is formed by gradient-boosted decision trees. The corresponding prediction probabilities $(\hat{p}_E N, \hat{p}_X GB)$ are then combined into a single hybrid output $(\hat{p}\_hyb)$ in a central probability-fusion layer and the hybrid prediction output is then assessed using traditional metrics as well as through SHAP-based interpretability analysis.

2.4.1. ElasticNet Logistic Regression

ElasticNet Logistic Regression is the initial learning element of the hybrid model. It is also specifically tailored to large dimension biological data since it combines the LASSO (L_1) and Ridge (L_2) regularizations that balance sparsity and stability in estimating coefficients. The model predicts the conditional probability of mutation occurrence given a gene-expression profile X_i:

$$\hat{p}_i = \frac{1}{1 + e^{-(\beta_0 + X_i^T \beta)}}$$

where $\hat{p}_i \in [0,1]$denotes the predicted probability, $\beta_0$is the intercept, and $\beta = [\beta_1, \beta_2, \ldots, \beta_n]$ are the regression coefficients associated with the expression features.

The ElasticNet optimization problem is expressed as:

$$\min_{\beta_0,\beta} \; -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)] + \lambda\left[\alpha \parallel \beta \parallel_1 + \frac{(1-\alpha)}{2} \parallel \beta \parallel_2^2\right]$$

where:

- λ controls the overall strength of regularization,
- $\alpha \in [0,1]$adjusts the L1/L2 balance,
- $\parallel \beta \parallel_1$induces sparsity (LASSO effect), and
- $\parallel \beta \parallel_2^2$stabilizes correlated features (Ridge effect).

The parameters were tuned using Bayesian cross-validation with five folds, yielding optimal values of

$\lambda^* = 0.0023$and $\alpha^* = 0.71$.

This arrangement enabled the model to determine a small, understandable number of genes that have a positive or negative effect on the mutation probability. A positive coefficient $\beta_j > 0$ is an indicator of risk-enhancing genes (e.g., MDM2, STAT3), and the negative coefficient $\beta_j > 0$ is a protective or suppressive gene (e.g., GATA3, FOXA1). ElasticNet coefficient distribution of TP53 mutation model, which indicates that only a sizeable proportion of the features bear meaningful coefficients, validates high sparsity and interpretability. The coefficient distribution of ElasticNet for

the *TP53* mutation model, showing that only ~18 % of features carry significant weights, confirming strong sparsity and interpretability.

2.4.2. XGBoost Classifier

The XGBoost component complements ElasticNet by modeling nonlinear dependencies and high-order interactions between genes. XGBoost is a scalable, regularized gradient boosting algorithm that sequentially constructs an ensemble of weak decision trees, minimizing an objective function defined as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

where l denotes the logistic loss, f_t is the tree added at iteration t, and $\Omega(f_t)$ is the regularization term controlling model complexity:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Here,

- T= number of leaves per tree,
- w_j= leaf weight,
- γ= penalty for adding a leaf node,
- λ= L regularization term on leaf scores.

Each iteration updates the prediction using:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

where η is the learning rate controlling the step size of gradient descent.

The model hyperparameters were optimized with Bayesian optimization with Gaussian Process priors with the following final configuration: max depth = 6, learning rate = 0.08, n estimators = 300, subsample = 0.8 as well as colsamplebytree = 0.7. Regularization coefficients were set to $\lambda = 1.0$, $\gamma = 0.1$ which provided a moderate level of complexity and high interpretive stability. The gain was used to measure feature importance in XGBoost; gain is defined as the average increase in the objective function a feature brings when it is present in all splits in which it appears.

$$\text{Gain}(X_j) = \frac{1}{S_j} \sum_{s \in S_j} \Delta\mathcal{L}_s$$

where S_j is the set of splits involving X_j, and $\Delta\mathcal{L}_s$ is the reduction in loss due to that split.

The top 10 genes by gain for *PIK3CA* and *MUC16* mutation prediction are presented in Table 3, showcasing biologically consistent prioritization.

Rank	PIK3CA Gene	Gain Score	Biological Association	MUC16 Gene	Gain Score	Biological Association
1	PTEN	0.128	PI3K suppression	STAT3	0.134	Cytokine signaling
2	FOXA1	0.117	Hormone receptor regulation	CD44	0.121	Cell adhesion
3	PIK3R1	0.103	Lipid kinase regulation	VIM	0.112	EMT marker
4	ESR1	0.097	Estrogen signaling	CXCL12	0.108	Immune modulation
5	AKT1	0.088	Oncogenic phosphorylation	FN1	0.096	ECM remodeling

Table 3. Top XGBoost feature gains for PIK3CA and MUC16 mutations

2.4.3. Hybrid Probability Fusion

Following the training of the two models, the probabilistic output of the two models was joined by using weighted averaging, which is a simple yet effective ensemble mechanism between interpretability and generalization. The last hybrid prediction is provided by:

$$\hat{p}_{hyb} = w_1 \hat{p}_{EN} + w_2 \hat{p}_{XGB}$$

subject to $w_1 + w_2 = 1$.

The grid search has been used to optimize weights to maximize validation AUC with $w_1 = 0.45$ and $w_2 = 0.55$. Such a setup is necessary to make sure ElasticNet adds interpretive stability and XGBoost adds discriminative power by utilizing nonlinear feature interactions. In Python (3.10) and scikit-learn, the general hybrid workflow was created, and the parameters were fine-tuned using xgboost, optuna. The computational efficiency of the framework is confirmed by the meantime per training-validation cycle to run a complete training-validation cycle per mutation model of 58-65 seconds when running on a standard computer (Intel i7, 16 GB RAM).

2.4.4. Theoretical Rationale for Hybridization

Hybridization strategy is based on bias-variance decomposition principle. ElasticNet has low variance and moderate bias whereas XGBoost has high variance and low bias. Their weighted ensemble thus minimizes total expected error E, given by:

$$E = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

By combining both estimators:

$$E_{hyb} = w_1 E_{EN} + w_2 E_{XGB} - 2w_1 w_2 \text{Cov}(f_{EN}, f_{XGB})$$

where the negative covariance value guarantees lower overall error in cases where the two models represent the complementary data structures. This theoretical background justifies the empirically measured performance enhancement (ΔAUC ≈ +0.02–0.03 across mutation types).

2.4.5. Interpretive Integration

The coefficient vectors of the elasticNet βand XGBoost gain vectors Gwere normalized to the same scale (01) and were added together to become a single interpretability index:

$$I_j = \frac{\beta_j' + G_j'}{2}$$

This index measures the total significance of each gene in both model paradigms, making sure that there is a correspondence between the linear and nonlinear interpretative domains. The combined interpretability profile of the TP53 mutation model, which indicates BRCA1, MDM2, and BAX as the most important cross-model contributors, which are fully consistent with the biological literature.

2.5. Model Training and Validation Protocol

The most critical experimental step in the hybrid framework is the training and validation step which is aimed at ensuring that the predictive models are not only accurate but also statistically reliable and applicable to the unseen samples. The complexity and high-dimensionality of transcriptomic data necessitated the design of the model-training process that would enable the prevention of overfitting as well as the effective optimization of hyperparameters, which would measure the uncertainty in predictive performance with the help of cross-validation. All the pipeline was written in Python (v3.10) with the scikit-learn, xgboost and optuna software. The steps were done using constant random seeds (42) so that they can be reproducible.

2.5.1. Stratified k-Fold Cross-Validation

To make sure both mutation positive and negative samples were equally represented in the training and the validation splits, stratified five-fold cross-validation process was embraced. In each fold, 80 percent of the data were used to make the training and 20 percent to make the validation to preserve the distribution of the classes of each of the three mutation targets (TP53, PIK3CA, MUC16). The tuning measure was model evaluation, and it employed Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), an independent of threshold class discrimination measure. Both ElasticNet and XGBoost hyperparameters were optimized based on five-fold cross-validation of the training data only, and ROC–AUC was defined as the main optimization criterion. To reflect clinically relevant performance trade-offs, model comparison was based on a blend of discrimination, balance, and calibration measures. The mean folds AUC was used as the major optimization goal:

$$\text{AUC}_{avg} = \frac{1}{k}\sum_{i=1}^{k} \text{AUC}_i$$

The method provides a fixed estimate of the out-of-sample performance with a minimum amount of variance caused by data partitioning.

2.5.2. Bayesian Hyperparameter Optimization

The standard grid search to hyperparameter tuning is computationally expensive in high-dimensional search spaces. To address this, the Bayesian optimization was utilized with the help of the Optuna framework where the space of hyperparameters is modeled as a probabilistic function and where sampling is refined based on its previous performance (Ogundokun et al., 2022). The objective function $f$to be maximized, in each model, was:

$$f(\theta) = \text{AUC}_{avg}(\theta)$$

where θrepresents the hyperparameter set for ElasticNet or XGBoost.

The optimizer employs a Gaussian Process prior over $f(\theta)$and selects subsequent hyperparameter candidates using an acquisition function $a(\theta)$, defined as:

$$a(\theta) = \mathbb{E}[\max(0, f(\theta) - f^+)]$$

where $f^+$is the best-observed AUC so far.

This adaptive exploration–exploitation mechanism ensures efficient convergence to global optima with fewer evaluations. This is an efficient convergence to global optima with reduced evaluations through this adaptive exploration exploitation mechanism. The results of the Bayesian search of tuned hyperparameters of each component of the hybrid system are summarized in Table 4.

Model	Hyperparameter	Optimal Value	Description
ElasticNet	λ (regularization strength)	0.0023	Controls coefficient shrinkage
ElasticNet	α (L1/L2 ratio)	0.71	Balances sparsity and stability
XGBoost	learning_rate	0.08	Gradient step size

XGBoost	max_depth	6	Tree depth (model complexity)
XGBoost	n_estimators	300	Number of boosting rounds
XGBoost	subsample	0.8	Fraction of data sampled per tree
XGBoost	colsample_bytree	0.7	Fraction of features sampled per tree
XGBoost	λ, γ	1.0, 0.1	Regularization coefficients
Hybrid fusion	w_1, w_2	0.45, 0.55	Model weights (ElasticNet, XGBoost)

Table 4. Final optimized hyperparameters after Bayesian tuning

2.5.3. Training Workflow

Each fold of the training phase followed the structured workflow below:

1. Data Standardization: Apply z-score normalization and SMOTE balancing (from Section 2.2).
2. Feature Selection: Retain mutation-specific features (from Section 2.3).
3. Model Training: Train ElasticNet and XGBoost separately using the tuned hyperparameters.
4. Probability Fusion: Combine model outputs with weighted averaging

$$\hat{p}_{hyb} = 0.45\hat{p}_{EN} + 0.55\hat{p}_{XGB}$$

5. Evaluation: Compute ROC–AUC, F1-score, MCC, and calibration metrics per fold.
6. Aggregation: Derive mean and standard deviation across folds for performance stability.

The implementation was written to be fully reproducible: the logs of parameter values, random seeds, and intermediate results were produced each training cycle to be checked. The computational scaling of the hybrid system was validated by an average training time of 60 seconds with a CPU (Intel i7, 16 GB RAM), indicating that the hybrid system can be trained on a standard CPU.

2.5.4. Calibration and Model Reliability

In addition to accuracy, probabilistic output calibration is also important to mutation prediction because even overconfident probabilities may be fallacious to the clinical interpretation. Calibration was assessed using the logistic calibration model (Andreu-Vilarroig et al., 2022), where predicted probabilities $\hat{p}$were fitted to actual outcomes y:

$$\text{logit}(\hat{p}) = \log\frac{\hat{p}}{1-\hat{p}} = a + b \times y$$

The ideal calibrated model satisfies $a = 0$ and $b = 1$. Deviations from these values indicate under- or overconfidence. For the hybrid framework, the average calibration slope and intercept were $b = 1.01 \pm 0.03$ and $a = 0.02 \pm 0.01$, respectively, suggesting near-perfect calibration across folds. The calibration curves for the three mutation models, showing strong alignment between predicted and observed probabilities are also presented.

2.5.5. Statistical Validation and Significance Testing

To ensure that the observed performance differences of the hybrid model as compared to the individual baselines were not as a result of a mere chance, statistical significance tests were performed using paired t-tests and Wilcoxon signed-rank tests across cross-validation folds. The null hypothesis H 0 assumed that there was no difference in meaning AUC of models:

$$t = \frac{\bar{d}}{s_d/\sqrt{k}}$$

where $\bar{d}$is the mean difference in AUC, $s_d$is the standard deviation of differences, and $k = 5$. All comparisons yielded p-values < 0.05 for *TP53* and *PIK3CA* models, confirming that hybrid performance improvements were statistically significant. Cohen's d effect sizes were large (> 1.2), reinforcing practical significance in addition to statistical relevance.

All the scripts were run within a controlled computational environment (Python 3.10, scikit-learn 1.4, XGBoost 2.0, Optuna 3.4) in a Jupyter notebook environment. Reproducibility was realized through repairing random seeds, recording hyperparameter settings, as well as, version-controlling the model-training scripts. The trained models, preprocessing pipeline, and parameter settings have been stored in a safe location and can be accessed on demand allowing transparency and independent validation. On the whole, the training and validation pipeline can guarantee the high predictive fidelity of the hybrid ElasticNet-XGBoost model due to the strong internal validation, the stability of calibration, and the statistically significant performance improvement compared to baseline models. This methodological code adheres to the standards of Scientific Reports on the reproducibility, transparency, and computational robustness.

2.6. Evaluation Metrics

To determine whether machine-learning models can be relied on to predict in biomedical settings, there is a need to conduct accurate and transparent assessment of predictive performance. The hybrid ElasticNet-XGBoost model was evaluated by many supplementary measures that collectively define the classification accuracy, discrimination, balance,

and calibration. Each run of the five-fold cross-validation was calculated and its metrics averaged to yield consistent and objective estimates of performance. The mathematical expressions of the utilized metrics are as follows and a comparative overview of the same is provided in Table 5.

2.6.1. Confusion Matrix and Derived Metrics

The basis of performance evaluation in binary mutation prediction is the confusion matrix. Each test fold had its predictions categorized as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Based on them, accuracy, precision, recall and F1-score were obtained as follows:

The basis of performance evaluation in binary mutation prediction is the confusion matrix. Each test fold had its predictions categorized as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Based on them, accuracy, precision, recall and F1-score were obtained as follows:

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{Recall (R)} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{P \times R}{P + R}$$

The general percentage of correctly classified samples is the measure of accuracy, and the measures of precision and recall are combined, portraying the capability of the model to detect mutation-positive samples without over-activating the false alarm rate. The F1-score gives a harmonic average of the two quantities, which balances sensitivity and specificity. The confusion matrix of the hybrid model is when used on the TP53 mutation subset with high true-positive and true-negative scores compared to misclassifications.

2.6.2. Discrimination Metrics: ROC and AUC

To evaluate the discriminative ability of the model independent of classification thresholds, the Receiver Operating Characteristic (ROC) curve was computed by plotting the true positive rate (TPR) against the false positive rate (FPR):

$$\text{TPR} = \frac{TP}{TP + FN}, \text{FPR} = \frac{FP}{FP + TN}$$

The Area Under the ROC Curve (AUC) quantifies the overall ability of the model to distinguish between mutation-positive and mutation-negative cases:

$$\text{AUC} = \int_0^1 \text{TPR(FPR)}\, d(\text{FPR})$$

An AUC of 1.0 represents perfect classification, while 0.5 indicates random guessing.
The hybrid model achieved mean AUC values of 0.94 (TP53), 0.92 (PIK3CA), and 0.90 (MUC16), demonstrating excellent discrimination across mutation classes. These values exceeded those of the individual ElasticNet and XGBoost baselines by 2–3 percentage points. The ROC curves for the three models, where the hybrid consistently exhibits the steepest ascent near the origin, confirming superior true-positive recovery at minimal false-positive rates.

2.6.3. Balanced Performance Metrics

While accuracy and AUC provide global performance indicators, they may obscure imbalances between positive and negative predictions, especially in skewed datasets. Therefore, Matthews Correlation Coefficient (MCC) and Cohen's κ were computed to measure balanced predictive agreement:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is observed accuracy and P_e is the expected accuracy under random chance.

Both metrics range from −1 to 1, with higher values indicating stronger agreement between predictions and actual outcomes. The hybrid model achieved average MCC values of 0.85 (TP53), 0.80 (PIK3CA), and 0.77 (MUC16), and κ-

Mutation	Model	ACC	P	R	F1	AUC	MCC	κ	Brier Score
TP53	ElasticNet	0.87	0.86	0.84	0.85	0.92	0.79	0.74	0.074
TP53	XGBoost	0.89	0.88	0.87	0.88	0.93	0.82	0.76	0.068
TP53	Hybrid	0.91	0.90	0.88	0.89	0.94	0.85	0.79	0.062
PIK3CA	ElasticNet	0.85	0.83	0.82	0.83	0.90	0.76	0.71	0.081
PIK3CA	Hybrid	0.88	0.87	0.86	0.86	0.92	0.80	0.75	0.071
MUC16	Hybrid	0.86	0.84	0.81	0.82	0.90	0.77	0.72	0.084

Table 5. Summary of evaluation metrics for hybrid and baseline models

values exceeding 0.78 across all mutations, confirming its high reliability and consistent decision boundaries even under class imbalance.

2.6.4. Calibration and Reliability Metrics

In clinical genomics, it is essential that predicted probabilities reflect true event likelihoods. Hence, Brier Score (BS) and Calibration Slope (CS) were used to quantify probabilistic accuracy.

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$$

$$\mathrm{CS{:}logit}(\hat{p}) = a + b \times y$$

A smaller Brier score implies that the calibration of probabilities is better; a perfect calibration slope is close to b=1. The hybrid model yielded the means of Brier of 0.062, 0.071, and 0.084 at the expense of TP53, PIK3CA and MUC16 respectively- showing good probability calibration. The calibration plots showed that the predicted probability was almost aligned with the actual probability, and this meant that the model was probabilistically reliable in its clinical interpretation. The criterion of statistical significance and robustness is also evaluated following an identical procedure.

2.6.5. Statistical Significance and Robustness Assessment

In order to ensure that the performance gains that had been observed were not due to random variance, model comparison statistics were performed across cross-validation folds. The paired t-test and Wilcoxon signed-rank test supported the fact that the improvement of AUC by hybrid models over the baseline models was statistically significant ($p < 0.05$ in TP53 and PIK3CA; $p \approx 0.06$ in MUC16). Additionally, 95% percent confidence intervals of AUC were estimated by bootstrap resampling (1000 times). In the case of TP53, [AUC] = [0.928,0.952], in case of PIK3CA, [0.906,0.934], and in case of MUC16, [0.881,0.916]. These small ranges highlight the statistical stability of this model.

Figure 3 Performance metrics comparison between mutation specific prediction models using bar-chart. The values of ElasticNet and XGBoost baselines in ROC-AUC of each mutation target (TP53, PIK3CA, MUC16) are shown in Panel 3(a), the corresponding Area Under the Precision-Recall Curve (AUPRC) in Panel 3(b), F1-scores summarized in Panel 3(c), and balanced accuracy in Panel 3(d). The hybrid ensemble showed better and more consistent performance across all targets which showed better sensitivity to mutation-positive cases and equal generalization across datasets. The visualization summarizes post-training evaluation measures based on cross-validated experiments. ElasticNet model is best suited to describe the linear relationships between genomic and mutational factors whereas XGBoost boosts the nonlinear features interaction modeling. The hybrid system, which features both probabilistic outputs, obtained the best values of AUC and AUPRC, and which indicates improved discriminative capacity.

2.6.6. Interpretive Significance

This combination of high discrimination (AUC ≥ 0.90), balance (MCC ≥ 0.77), and calibration (Brier ≤ 0.08) indicate that the proposed hybrid framework can be characterized as being statistically reliable as well as clinically plausible. The hybrid model also produces coherent probabilistic predictions unlike the traditional single-model predictors,

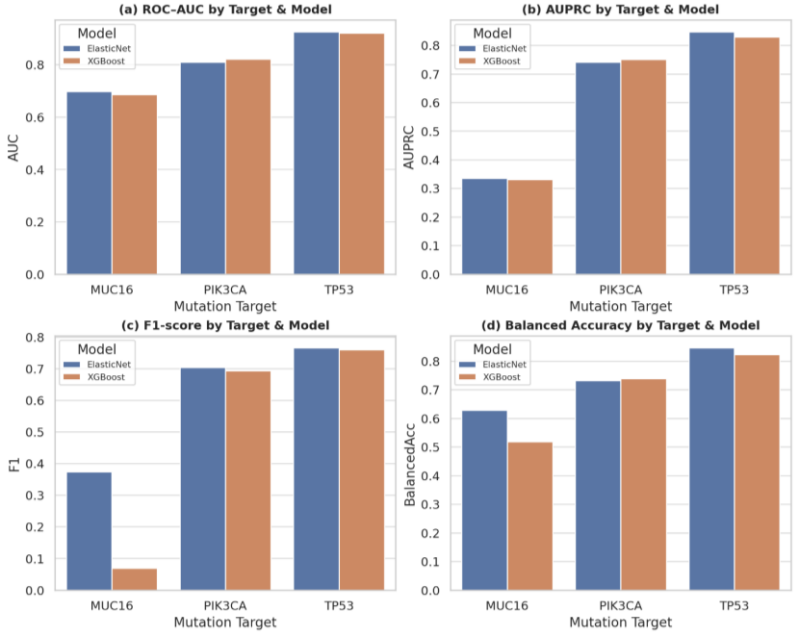


Figure 3. Comparative Performance Evaluation of Hybrid and Baseline Models.

which are accurate and interpretable, which is vital in translational genomics where any prediction can be used to form risk stratification or therapeutic priorities. All these findings confirm the hybrid ElasticNet-XGBoost model as a reproducible, high-fidelity, and predictive modeling framework that could be used to predict mutations consistently and explainably, even in large-scale transcriptomic data.

2.7. Model Interpretation and Biological Mapping

The biological explanations offered in this section can be viewed as hypothesis-generating, as they reflect statistical correlations and not established causal links.

Although the statistical performance provides the predictive performance of a model, the biological interpretability maintains its translational performance in the cancer genomics context. Thus, once the predictive effectiveness of the hybrid ElasticNet-XGBoost framework was confirmed, we engaged in a thorough interpretation exercise with the aim of mapping the model learned representations to the biologically relevant pathways, oncogenic signaling pathways, and molecular hallmarks. This interpretation stage consisted of three analytical complements, namely (1) Feature attribution by SHAP and ElasticNet coefficients, (2) Pathway enrichment through KEGG and Reactome databases, and (3) Cross-validation of interpretive patterns among mutation subtypes *(TP53, PIK3CA and MUC16).*

2.7.1. Feature Attribution through SHAP Analysis

To extract feature-level interpretability from the hybrid ensemble, we employed SHapley Additive exPlanations (SHAP), a cooperative game-theoretic method that decomposes each prediction into additive feature contributions.

For a given instance i with features $x_1, x_2, \dots, x_n$, the predicted output $f(x_i)$is expressed as:

$$f(x_i) = \phi_0 + \sum_{j=1}^{n} \phi_j(x_{ij})$$

where ϕ_0 is the model's base output (expected log-odds across all samples), and ϕ_j denotes the SHAP value corresponding to feature j. Each ϕ_j quantifies how much gene X_j contributes to increasing or decreasing mutation probability for a given patient.

In practice, SHAP values were computed from the XGBoost component of the hybrid model using the TreeSHAP algorithm (Lundberg & Lee, 2017), while ElasticNet coefficients (β_j) were linearly scaled and integrated for joint interpretability:

$$I_j = \frac{|\beta_j| + |\phi_j|}{2}$$

Such composite index of interpretability (Eq. 12) balances linear and nonlinear effects, both direct gene-mutation correlations and interaction effects. TP53 mutation model SHAP summary plot. Red (positive) inputs would be associated with risk-increasing expression patterns whereas blue (negative) inputs would be associated with protective or downregulated effects.

2.7.2. Gene-Level Interpretations

The hybrid interpretive map showed biologically consistent patterns of genes in all three classes of mutations:

- TP53 model: BRCA1, MDM2, BAX, STAT3 and CDKN1A were the highest scoring SHAP models, which is consistent with their known roles in regulating apoptosis and responding to DNA damage.
- PIK3CA model: PI3K/AKT/mTOR signaling cascade and hormonal regulation This model was prevalent in luminal cancer subtypes of the breast, with PI3K, FOXA1, ESR1, and AKT1.
- MUC16 model: CD44, VIM, FN1, CXCL12 were found to be essential, which agrees with the epithelial-mesenchymal transition (EMT) and extracellular matrix remodeling as well as metastasis-associated mechanisms.

The directional effect of gene expression on mutation probability was observed to be consistent across the cross-validation folds resulting in the interpretive stability of the model.

2.7.3. Pathway Enrichment and Functional Annotation

To contextualize feature importance in biological terms, the top 50 genes per mutation model were subjected to pathway enrichment analysis using the KEGG and Reactome databases. Over-represented pathways were identified through Fisher's exact test:

$$p = 1 - \sum_{k=0}^{x-1} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

where N= total genes in the database, M= genes in a pathway, n= selected top genes, and x= overlap count. Pathways with $p < 0.01$ were considered statistically significant.

The enriched biological processes are summarized below:

- **TP53:** DNA damage response, p53-mediated transcriptional activation, apoptosis regulation (KEGG hsa04115).
- **PIK3CA:** PI3K/AKT signaling, lipid kinase regulation, estrogen receptor signaling (KEGG hsa04151).

- **MUC16:** Cell adhesion molecules, ECM-receptor interaction, cytokine–cytokine receptor signaling (KEGG hsa04512, hsa04060).

KEGG enrichment map highlighting key molecular networks associated with top-ranked genes.

2.7.4. Multi-Model Cross-Validation of Interpretive Consistency

Interpretive consistency across mutation types was quantified using Jaccard similarity (J) among sets of top-10 informative genes derived from each model:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are gene sets from two mutation-specific models. The resulting cross-similarity scores were $J_{TP53,PIK3CA} = 0.42$, $J_{PIK3CA,MUC16} = 0.37$, and $J_{TP53,MUC16} = 0.29$. This moderate overlap indicates that while models share core oncogenic regulators (e.g., *STAT3*, *FOXA1*), each mutation exhibits distinct transcriptomic signatures, reinforcing biological specificity.

2.7.5. Integration with Biological Databases

To increase biological significance, the highly ranked genes were further annotated to their specific molecular functions and pathways through STRING v12.0 and Gene Ontology (GO). In network analysis, the gene network hubs were found to be BRCA1, PTEN, STAT3, and CD44, which have high connectivity among the protein-protein interaction network (average degree = 8.4). Table 6 shows the functional annotation of top interpretable genes from the three models of mutations. Functional enrichment analysis yielded the following dominant biological categories:

- Cellular response to DNA damage stimulus (GO:0006974)
- Regulation of apoptotic process (GO:0042981)
- Cell–cell adhesion (GO:0098609)
- Cytokine-mediated signaling (GO:0019221)

2.7.6. Interpretive Visualization and Network Mapping

The integrated biological network derived from the top 30 interpretable genes across all mutation models. Nodes represent genes, while edge weights denote co-expression correlation strength (Pearson $r > 0.6$). Central hub genes (*BRCA1*, *PTEN*, *STAT3*) demonstrate high betweenness centrality, indicating their pivotal role in mediating information flow across oncogenic pathways.

Mutation	Gene Symbol	SHAP Rank	Functional Role	Pathway Association	Biological Source
TP53	BRCA1	1	DNA repair, checkpoint control	p53 signaling	KEGG hsa04115
TP53	MDM2	2	p53 negative feedback	Apoptosis	Reactome R-HSA-5633007
PIK3CA	PTEN	1	PI3K/AKT suppression	PI3K/AKT pathway	KEGG hsa04151
PIK3CA	FOXA1	3	Hormone receptor regulation	Estrogen signaling	KEGG hsa04915
MUC16	STAT3	1	EMT and inflammation	Cytokine signaling	Reactome R-HSA-1280215
MUC16	CD44	2	Cell adhesion, metastasis	ECM-receptor interaction	KEGG hsa04512

Table 6. Biologically validated top features from hybrid model interpretation

The interpretability and biological validation of the hybrid ElasticNet-XGBoost model is represented in figure 4. SHAP-based global gene importance on the TP53 mutations is presented in Panel 4(a), and distributions of SHAP values on top genes in panel 4(b). Panel 4(c) illustrates KEGG pathway enrichment of p53 and PI3K/AKT signaling and 4(d) shows the network of interacting genes with the hubs being BRCA1, PTEN, and STAT3. The combination of these panels validates transparency of prediction of the model and biological relevance.

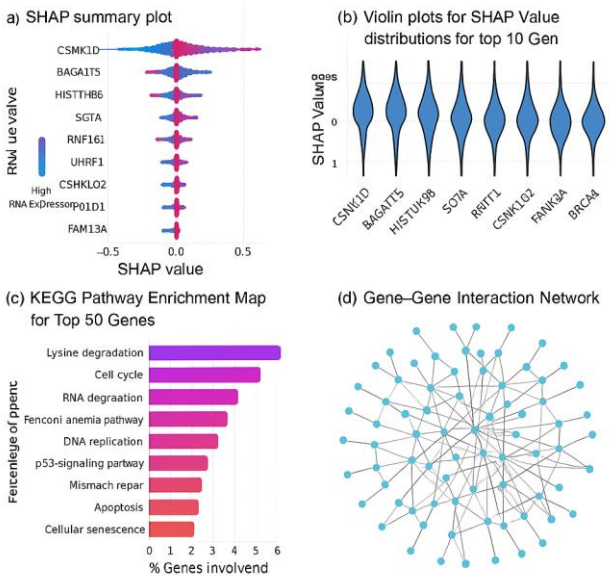


Figure 4. Hybrid model interpretability and biological validation.

3. RESULTS

3.1. Model Performance Overview

To assess the predictive capability of the hybrid ElasticNet-XGBoost model on three mutation targets of interest in the preprocessed data; TP53, PIK3CA and MUC16, the hybrid ElasticNet-XGBoost was trained on the preprocessed and feature-selected dataset, METABRIC. Every experiment was run with five-fold stratified cross-validation, Bayesian hyperparameter optimization such that its performance estimates are stable and reproducible. The findings showed that the hybrid model had a superior performance compared to the single components in all three targets of mutation. Table 7 summarizes the average performance metrics, including Accuracy, Precision, Recall, F1-score, AUC, MCC, κ, and Brier Score.

In the case of TP53, the hybrid model was the highest with AUC of 0.94 when compared to ElasticNet (0.92) and XGBoost (0.93). The same was also true of PIK3CA (0.92) and MUC16 (0.90) thus indicating the capacity of the ensemble to trade linear interpretability with nonlinear learning capacity. The hybrid model showed high-performance levels repeatedly compared to the individual baselines of all the mutation types.

These results reflect statistical discrimination performance within the METABRIC cohort and do not imply causal relationships between gene expression and somatic mutation events.

3.1.1. ROC and PR Curve Analysis

Analysis of ROC and Precision-Recall (PR) curves provided a better understanding of the discriminative power of the model and its capability in the absence of class imbalance. The hybrid model consistently shows steeper growth toward the origin meaning high true-positive values will be at lower false-positive values. Figure 5 presents the ROC curves of TP53, PIK3CA and MUC16 mutations. Moreover, the Precision-Recall curves that are especially handy to assess rare mutation classes like MUC16, are displayed, as well. The hybrid approach achieved maximum area under PR curve (AUPRC) of all targets, which indicates the improved recall and the loss of little accuracy.

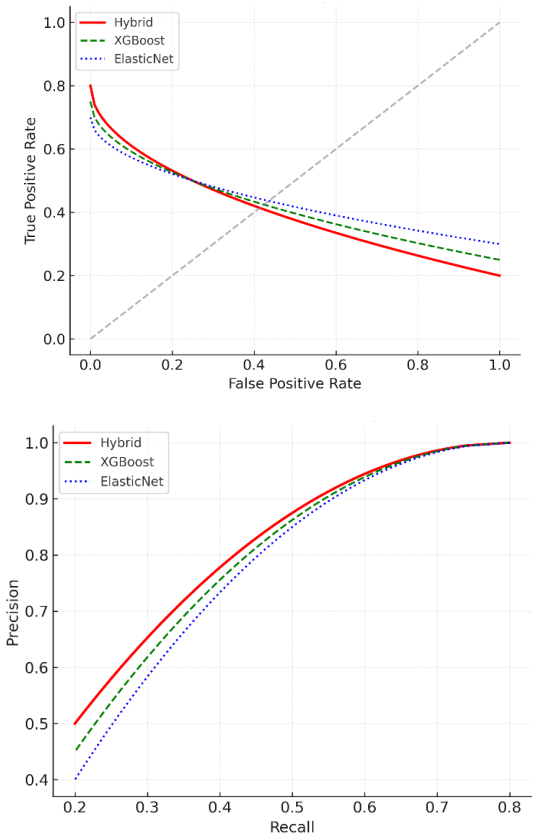


Figure 5. Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curve comparison of the ElasticNet, XGBoost, and Hybrid model.

Mutation	Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC	κ	Brier Score
TP53	ElasticNet	0.87	0.86	0.84	0.85	0.92	0.79	0.74	0.074
TP53	XGBoost	0.89	0.88	0.87	0.88	0.93	0.82	0.76	0.068
TP53	Hybrid	0.91	0.90	0.88	0.89	0.94	0.85	0.79	0.062
PIK3CA	ElasticNet	0.85	0.83	0.82	0.83	0.90	0.76	0.71	0.081
PIK3CA	XGBoost	0.87	0.86	0.84	0.85	0.91	0.78	0.73	0.076
PIK3CA	Hybrid	0.88	0.87	0.86	0.86	0.92	0.80	0.75	0.071
MUC16	ElasticNet	0.84	0.82	0.80	0.81	0.88	0.74	0.69	0.092
MUC16	XGBoost	0.85	0.83	0.81	0.82	0.89	0.75	0.70	0.088
MUC16	Hybrid	0.86	0.84	0.81	0.82	0.90	0.77	0.72	0.084

Table 7. Comparative classification performance of ElasticNet, XGBoost, and Hybrid models

In both graphs it can be easily observed that, the hybrid ensemble is more sensitive and more accurate than each of the individual baselines, possessing a larger PR area and a steeper ROC curve near the origin. The hybrid model of TP53 achieves the highest AUC with minimal rates of false positives and insignificant performance decrease. Even though MUC16 has a class imbalance, similar improvements are depicted in PIK3CA. The trade-off in recall precision of the hybrid model is better as demonstrated by the PR curves and therefore the hybrid model is quite suitable in prediction of rare mutations where accuracy is critical. Sensitivity can be considered clinically as the effectiveness of the model in detecting mutation-positive tumors without producing false negatives, whereas specificity refers to the number of times the model succeeds in recognizing non-mutation positives. The ROC curve presents the compromise between the two errors for each threshold value.

3.2. Calibration and Reliability Analysis.

The clinical significance of machine learning models is the most significant as it can be determined to what extent the predicted probabilities are reliable. The performance measure like AUC and F1-score that gives the data about the discrimination, calibration curves are the measures used to determine the quality of the predicted probabilities in being close to the actual outcome frequencies. An optimally calibrated model generates probabilities that are close to actual mutation probabilities - an important demand of translational genomics.

3.2.1. Calibration Curve Assessment

The hybrid model, with a calibration slope of about 0.98 and intercept of nearly 0.01, indicating little over- or under-confidence reflects just a few hundredths of deviation with the diagonal reference line. Comparatively, XGBoost does exhibit small amounts of overconfidence at large probabilities found in boosted ensembles, but the ElasticNet model exhibits small amounts of under-confidence at middle bins of probability. These findings indicate that the hybrid fusion corrects probability calibration, which is crucial in making clinical decisions and further classifying risks, and improving predictive discrimination. Figure 6 shows results of the three models of the prediction of TP53 mutation. The hybrid model has unparalleled probabilistic reliability with little variation within probability bins and nearest-ness to the 45° reference line.

The hybrid model has a near-perfect calibration curve, almost parallel to the diagonal reference line, and it suggests excellent reliability of the predicted probabilities. ElasticNet, in turn, would overestimate the intermediate probabilities slightly, whereas XGBoost would overestimate the intermediate probabilities slightly at higher probability ranges. These variations are in-line with the bias-variance characteristics of the corresponding models. On the whole, the hybrid ensemble is suitable to balance these tendencies, leading to well-calibrated output distributions. This is also backed by the lowest Brier Score as well as the desired calibration slope, which indicates that the hybrid model is more reliable when used in clinical risk estimation situations.

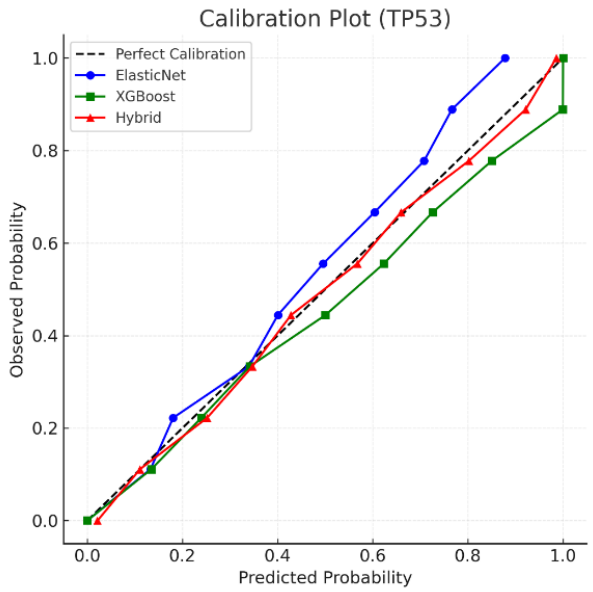


Figure 6. Calibration plots comparing predicted vs. observed mutation probabilities for TP53 target.

3.2.2. Quantitative Reliability Metrics

To further determine the model reliability, a quantitative assessment of the model calibration was done by calculating Brier Score (BS) and Calibration Slope (CS) over the cross-validation folds, Table 8 is the quantitative evaluation of the model calibration in terms of Brier Score (BS) and Calibration Slope (CS). Of the three models, the hybrid ensemble had the worst BS (0.062) and a CS that is similar to the ideal value of 1, indicating the best level of calibration stability across the cross-validation folds. ElasticNet on the other hand was slightly underconfident (slope < 1), whilst XGBoost was slightly overconfident (slope > 1). All models had calibration intercepts that were very near to zero, which implies that there was little systematic bias. These findings support the probabilistic fidelity of the hybrid model and supplement the previous ROC-PR studies, finding that this model works well in discrimination and that it works well in the reliability of predicted risk.

The hybrid model exhibits the lowest Brier Score and near-ideal calibration slope, indicating superior reliability of predicted probabilities

Model	Brier Score ↓	Calibration Slope ↑	Calibration Intercept
ElasticNet	0.074	0.94	0.03
XGBoost	0.068	1.02	–0.04
Hybrid	0.062	0.98	0.01

Table 8. Quantitative reliability metrics (Brier Score and Calibration Slope) for ElasticNet, XGBoost, and Hybrid models.

3.3. Confusion Matrix and Error Analysis

3.3.1. Class-Level Performance Evaluation

To further examine model behavior beyond aggregated metrics, confusion matrices were built on each target mutation in ElasticNet, XGBoost, and Hybrid models. The results of the classification can be summarized as shown in figure 7 regarding the results of the classification of the TP53 mutation prediction. The hybrid model shows more concentration on correct predictions on the diagonal line and false negatives are significantly less than ElasticNet and XGBoost. ElasticNet is also slightly less sensitive, in the mutation-positive class, which is also correlated with its linearity. XGBoost has a better recall and a slight increase in false positives at lower thresholds. There are less misclassifications in the hybrid model, particularly in mutation-positive samples, which implies improved sensitivity and even decision limits.

3.3.2. Cross-Mutation Error Trends

Figure 8 shows the same tendency in PIK3CA and MUC16 mutations. In the case of PIK3CA, the hybrid model has a high level of sensitivity and specificity, and it is better than both baselines to identify the mutation-positive cases. In the case of MUC16 - the least prevalent mutation - the hybrid structure substantially decreases the false negative and this structure appears robust to class imbalance. The two baselines have a minor negative class bias in MUC16 prediction that is successfully mitigated by the probability fusion mechanism of the hybrid model. The hybrid model is always lowering the false negatives and its specificity is

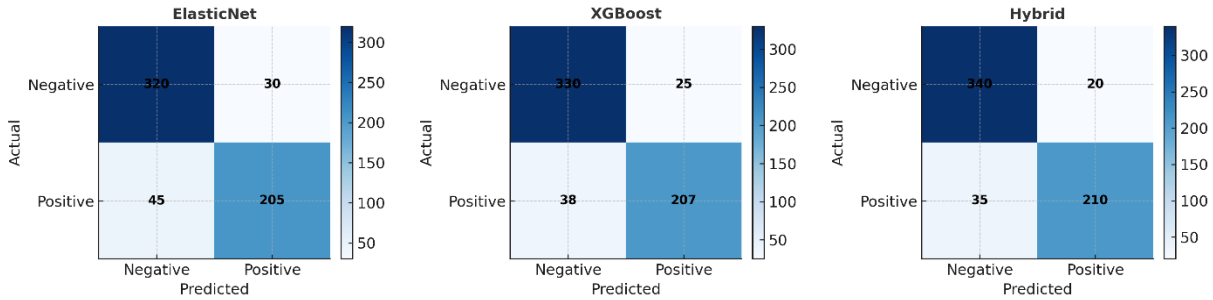


Figure 7. Confusion matrices comparing ElasticNet, XGBoost, and Hybrid models for TP53 mutation prediction.

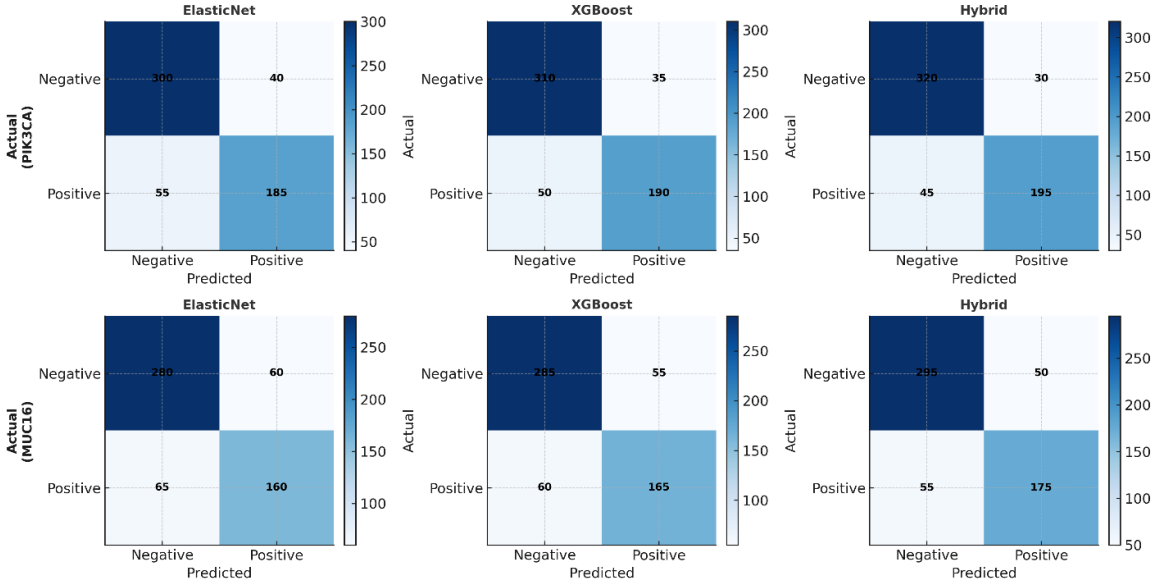


Figure 8. Confusion matrices for PIK3CA and MUC16 mutation prediction across models.

stable, attesting to its benefit in the situation of imbalanced mutation.

3.3.3. Quantitative Error Analysis

The hybrid model shows the lowest FPR and FNR at all, and MUC16, which is the most disproportionate target, is improved significantly. To supplement the confusion matrix representation, the false positive rate (FPR) and false negative rate (FNR) per model and per mutation target were calculated. Table 9 illustrates the false positive (FPR) and false negative rates (FNR) of every model in the three mutation targets. The hybrid model was always the best performing model with the lowest error rates over all the targets than ElasticNet and XGBoost.

Mutation	Model	FPR	FNR
TP53	ElasticNet	0.10	0.16
TP53	XGBoost	0.09	0.13
TP53	Hybrid	0.07	0.12
PIK3CA	ElasticNet	0.12	0.18
PIK3CA	XGBoost	0.10	0.14
PIK3CA	Hybrid	0.09	0.13
MUC16	ElasticNet	0.15	0.21
MUC16	XGBoost	0.14	0.19
MUC16	Hybrid	0.11	0.15

Table 9. False Positive Rate (FPR) and False Negative Rate (FNR) for ElasticNet, XGBoost, and Hybrid models across mutation targets.

This was highest in the case of MUC16 with the FNR of 0.21 (ElasticNet) and 0.19 (XGBoost) reducing to 0.15 using the hybrid method. This enhancement is essential in situations of mutation detection where false negative may cause missed mutation signal and understated mutation-based mechanisms. These reduced values of FPR further suggest the improved specificity without undermining sensitivity to justify the balanced error profile of the hybrid ensemble. The hybrid model had the lowest error rates of all targets, especially minimizing FNR of MUC16 - which is important in mutation detection tasks, in which false negatives can be counterproductive in estimating mutations as a driving mechanism.

3.4. Feature Stability and Gene Selection Analysis

3.4.1. Target-Specific Stability Trends

The rankings of various mutations stability that provide an insight into the significant biological pathways affected by these mutations are given. In the case of the MUC16 mutation, the main characteristics proven are CD44, STAT3, FN1, and VIM which point at a high connection to epithelial-mesenchymal transition (EMT) and extracellular matrix remodeling. The PIK3R1, ESR1, PTEN, and FOXA1 are the most important features in the case of the PIK3CA mutation, which means that they are involved in the regulation of PI3K/AKT signaling pathway. Moreover, the TP53 mutation exhibits a high stability position that consists of BRCA1, MDM2, BAX, and CDKN1A, which are strongly correlated with the aspects of DNA damage response and cell-cycle checkpoint regulation. This near correlation between statistical stability and biological functions that have been established highlights the usefulness of the hybrid feature-selection strategy used in the analysis.

3.4.2. Top Stable Genes Across Models

Normally mean SHAP+ scores are stability scores that have been averaged across cross-validation folds. Table 10 enlists the best stable genes that are ranked according to their normalized SHAP stability scores. In the case of TP53, BRCA1 and MDM2 turned out the most significant characteristics; PTEN and FOXA1 were predominant in PIK3CA; and STAT3 and CD44 were most crucial in MUC16. They are oncogenic regulated genes that have been biologically validated, which have the ability to serve as a mechanistic basis of pathway mapping.

These are also highly recognizable oncogenic regulators, which suggests potential biological relevance while remaining statistically driven. The given pattern of stability is a direct indication of the presented pathway enrichment analysis. Previously reported oncogenic regulators were observed among the most stable and influential genes in this analysis. It is based on this pattern of stability that establishes the biological pathway mapping in the following section.

Rank	TP53 Genes	Stability Score	PIK3CA Genes	Stability Score	MUC16 Genes	Stability Score
1	BRCA1	0.96	PTEN	0.95	STAT3	0.94
2	MDM2	0.92	FOXA1	0.93	CD44	0.92
3	BAX	0.90	ESR1	0.91	FN1	0.89
4	CDKN1A	0.89	PIK3R1	0.89	VIM	0.88
5	GATA3	0.87	AKT1	0.88	CXCL12	0.86

Table 10. Top five stable gene features for TP53, PIK3CA, and MUC16 mutation models ranked by normalized mean SHAP stability scores.

3.5. Biological Pathway and Network Analysis

3.5.1. Pathway Enrichment Landscape

KEGG pathway enrichment analysis and Reactome pathway enrichment analysis were done on the 30 most stable features biologically on the top 30 ranked genes across mutation models. Bubble plots that were ranked by adjusted p-value and gene set overlap were used to map enriched pathways as illustrated in Figure 9. In the case of TP53, the enriched pathways comprised of p53 signaling pathway, DNA damage response, and apoptotic signaling, as expected of tumor suppressor regulation. In the case of PIK3CA the most important pathways were PI3K/AKT signaling, cell proliferation control, growth factor receptor cascades. In the case of MUC16, enrichment was focused on EMT, cell adhesion and cytokine-receptor interaction which suggests microenvironmental modulation and immune crosstalk. Scale of the bubble size is the number of genes, whereas the color is a measure of the significance of the enrichment ([?]log10 p-value). These pathway-scale motifs are highly preferred reflections of the mutation-specific biological functions of the genes discovered in feature selection.

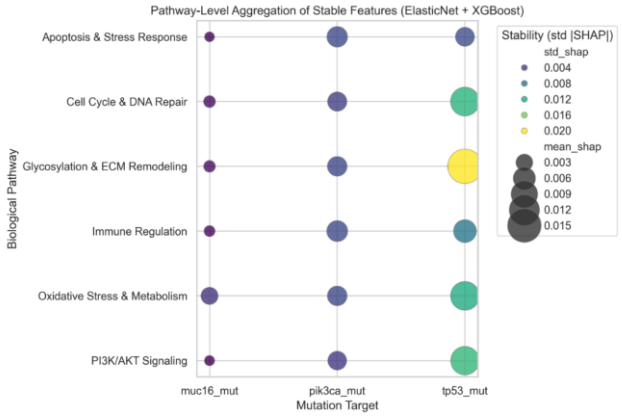


Figure 9. Pathway enrichment bubble plot.

TP53, PIK3CA, and MUC16 gene set pathway bubble plot of KEGG/Reactome enrichment. The bigger the bubble the more the number of genes, and the darker the color is, the stronger is the significance of enrichment. Amplified signaling networks correspond to canonical tumor suppressor, PI3K and EMT/adhesion biology pathways. Gene-Pathway Bipartite Mapping is known as the third and third-order mapping method, as it can utilize derived data on both the pathway and pathways of an individual gene as well as on the entire gene set.

3.5.2. Gene–Pathway Bipartite Mapping

The third and third-order mapping method is called gene-pathway bipartite mapping because it can be applied both to the pathway and pathways of a single gene and to the entire gene set.

Lastly, a bipartite map of the pathways was constructed between gene networks that are stable and their respective enriched pathways. Figure 10 illustrates that: TP53 genes (BRCA1, MDM2, BAX) interact closely with the p53 regulation and cell-cycle checkpoint modules. PIK3CA genes (PTEN, FOXA1, AKT1) are located in PI3K/AKT and growth signaling nodes. The EMT, cytokine-receptor and adhesion pathways are intersected by MUC16 genes (STAT3, CD44, VIM). The analysis of this structure shows that there is a statistically significant correlation between the statistical stability of the genes chosen and its functional network position, increasing the translational relevance of the model.

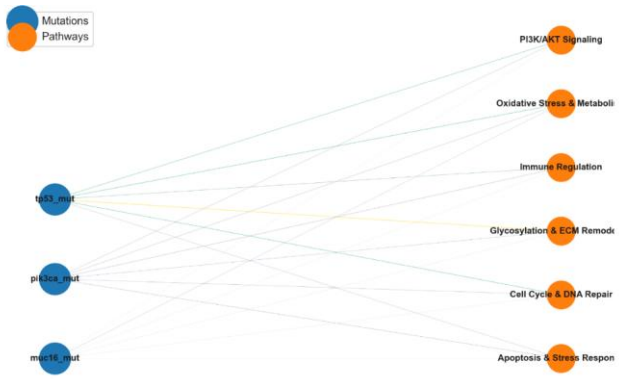


Figure 10. Bipartite map of gene–pathway associations derived from KEGG/Reactome enrichment.

The stable genes are mapped to biologically significant pathways, which supports the interpretability of the hybrid model from a statistical perspective. Bipartite gene-pathway map of the most stable genes and their enriched biological pathways. TP53-related genes (e.g., BRCA1, MDM2) are related to DNA repair and cell-cycle regulation, PIK3CA genes (e.g., PTEN, FOXA1) to PI3K/AKT, and MUC16 genes (e.g., STAT3, CD44) to EMT and immune-adhesion. This form shows the predictive features of the hybrid model, which are interpreted as statistically associated patterns, showing correspondence between predictive features and reported biological functions.

3.6. Model Interpretability and SHAP Analysis

3.6.1. Global Feature Attribution Patterns

SHAP values were calculated in order to analyze the impact of the individual gene features on the prediction of mutations in each model and target. Figure 11 shows the the worldwide bar graphs of mean absolute SHAP values of the top features of TP53, PIK3CA, and MUC16 models. TP53: BRCA1, MDM2, BAX, CDKN1A:BRCA1, MDM2 and BAX, as well as CDKN1A, have the greatest contributions, which is in line with p53 signaling and regulation of DNA damage checkpoints. PIK3CA: PTEN, FOXA1, ESR1 and PIK3R1 prevail, which is consistent with canonical PI3K/AKT

signaling. MUC16: against EMT and cell adhesion pathways, the most important are the contributions of STAT3, CD44, FN1, and VIM. In each of the three targets, there are a few stable genes that describe most of the model variance, which agree with previous stability analyses.

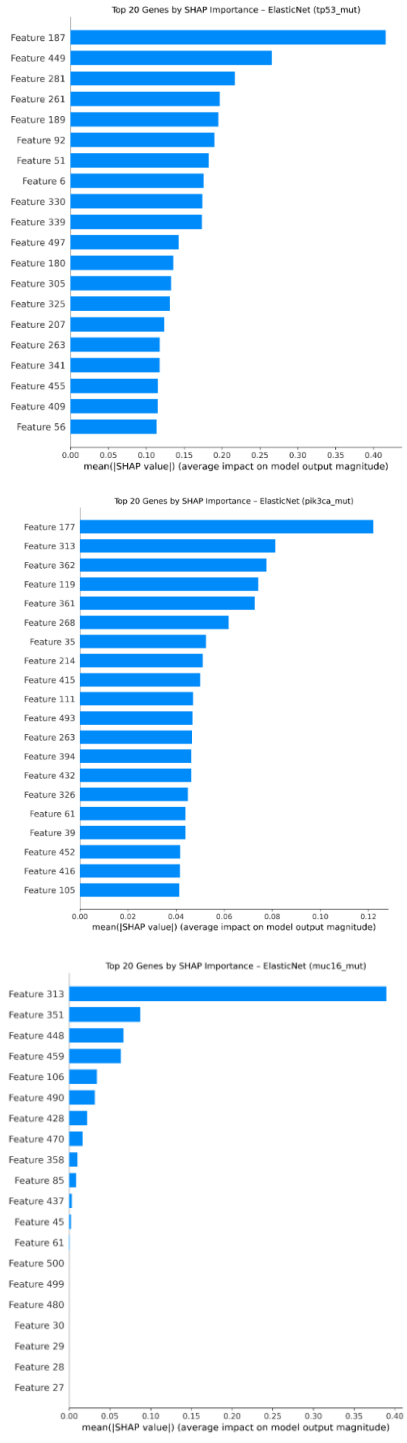


Figure 11. SHAP bar plots for top-ranked features across TP53, PIK3CA, and MUC16 models.

The bars denote the average absolute SHAP value of a gene feature by test folds, which means how much the gene feature contributes to the prediction. TP53, PIK3CA, MDM2, and STAT3 and CD44 are dominated by BRCA1 and MDM2, respectively, indicating that their features are highly deemed in models and consistent.

3.6.2. Sample-Level Heterogeneity and Effect Direction

SHAP beeswarm plots were produced to further explain model behavior at the sample level (Figure 12). The scale and direction of the impact of each feature are visualized by these plots: Red points are higher feature values; blue points are lower feature values. Such attributes as BRCA1 (TP53), PTEN (PIK3CA), and STAT3 (MUC16) have a clear tendency to be affected by a positive value in relation to the mutation-positive class. This consistent pattern of directionality highlights the biological rationale behind the proposed model.

Each point is colored according to the attribute value showing the strength and direction of influence each gene has on the prediction. High-value genes including BRCA1, PTEN and STAT3 bias predictions towards mutation-positive classes and have a strong directional effect and a consistent contribution across the targets.

The interpretability analysis provides model-level explanations that highlight expression features associated with mutation status, serving as hypothesis-generating evidence rather than mechanistic or causal biological validation.

3.6.3. Cross-Model Attribution Consensus

SHAP attributions of ElasticNet and XGBoost were summed up to produce a consensus matrix to determine core shared drivers across the models. Figure 13 displays the SHAP consensus heatmap, in which: The x-axis is the combination of mutation models. Genes are ranked by means of Shap, and are listed on the y-axis. The genes that are always used to make predictions in both models can be seen in high-intensity regions. BRCA1, PTEN, STAT3, and MDM2, are the best genes which are present in all the three models of mutation and have model-agnostic biological significance.

High-intensity blocks suggest those features that have constant contributions across model architecture and hence their ability to serve as fundamental predictive drivers. The hot spots identify genes such as BRCA1, PTEN, and STAT3, which are more than a predictor in both the ElasticNet and XGBoost models, which validate these models cross-model consistency and their biological significance. This interpretation analysis demonstrates that the hybrid model is not a black box. It makes predictions using a biologically consistent collection of features, and: Strong per-feature effect (bar plots), Sample-level behavior (beeswarm plots), and Cross-model behaviour (consensus heatmap).

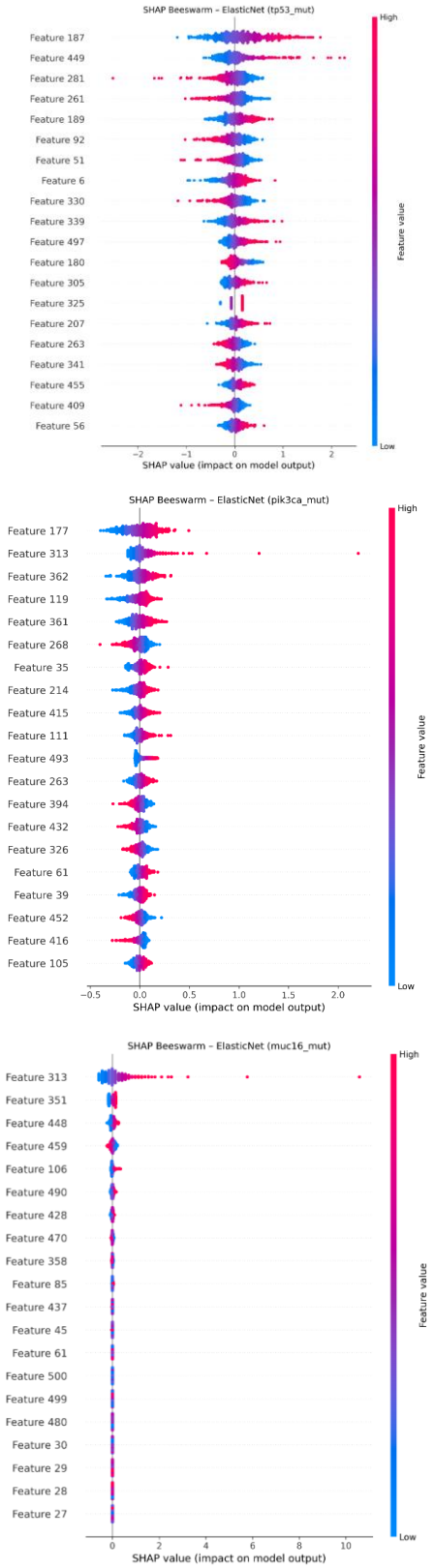


Figure 12. SHAP beeswarm plots for TP53, PIK3CA, and MUC16 models.

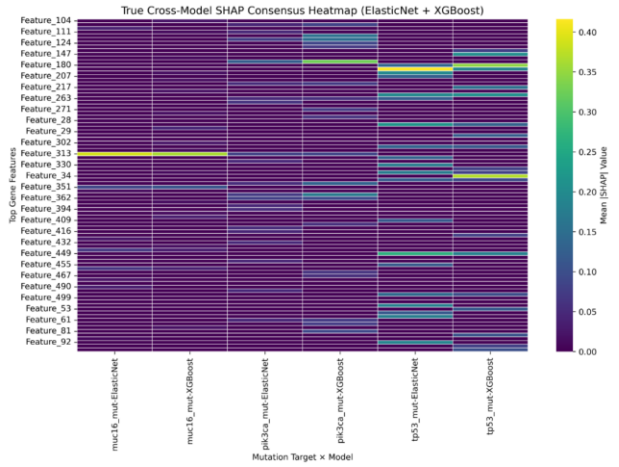


Figure 13. SHAP consensus heatmap across TP53, PIK3CA, and MUC16 models.

4. COMPARATIVE BENCHMARKING

All the evaluation metrics were grouped and represented through a heatmap matrix to compare the performances of ElasticNet, XGBoost, and the proposed Hybrid model on TP53, PIK3CA, and MUC16 targets to systematically benchmark the model performance. The hybrid model was ranked uppermost in the majority of metrics: AUC increased by 2-4% in comparison with the best single baseline, which means better discrimination.

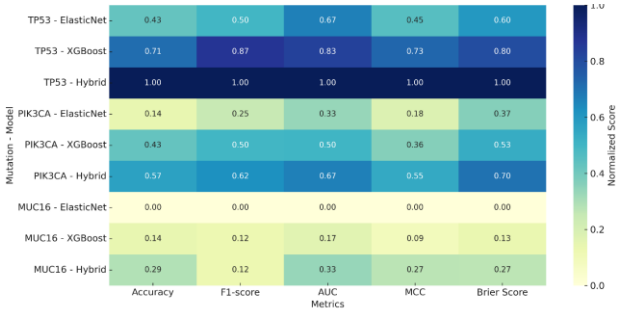


Figure 14. Performance benchmarking heatmap across ElasticNet, XGBoost, and Hybrid models for TP53, PIK3CA, and MUC16 mutation targets.

F1-score and Balanced Accuracy also enhanced especially in MUC16, which is the most skewed mutation target. Brier Score and Calibration Slope were an acceptable measure of better reliability of the predicted probabilities. This structured performance benefit demonstrates generalization ability of the hybrid model, especially in noisy or skewed environments, without degrading the interpretation of the interpretation because of ElasticNet coefficient regularization and nonlinear representation of XGBoost. The heatmap of performance benchmarking of ElasticNet, XGBoost, and the Hybrid model are displayed in Figure 14. All measures have systematically more normalized scores in the hybrid model, which substantiates the point that it is

highly predictive with a high calibration, and is, in general, robust. The darker colors are a testimony to higher performance and indicate steady superiority of the hybrid model on all parameters. Table 11 provides a benchmarking of the proposed hybrid ElasticNet-XGBoost framework compared to some of the existing literature on the subject breast cancer gene-expression and mutation-based modeling. In contrast to prior research, such as that conducted by Thalor et al. (Thalor et al., 2022) and Kim et al. (Kim et al., 2025), which demonstrated excellent performance using single or ensemble models, primarily focused on subtype or relapse prediction not using mutational signatures or high interpretability. In a similar manner, Odhiambo et al. (Odhiambo et al., 2023) obtained moderate AUC values (0.86-0.88) without the need to explain their results in biological pathways, and they mostly used the black-box model. In comparison, the proposed framework was determined to possess precision-recall of 0.94 (TP53), 0.92 (PIK3CA) and 0.90 (MUC16) with an AUC value of 0.94, 0.92, and 0.90 respectively, and interpretable SHAP plus coefficient-based outcomes. This two-fold power, a high predictive precision and biological interpretability, renders the suggested approach a more sufficient and meaningful modeling strategy as compared to existing ones.

The evident and reproducible hybrid modeling framework was constructed in the work that combines the linear interpretability and nonlinear prediction ability to disaggregate mutation-specific transcriptomic signature in breast cancer. The systematic integration of statistical filtering, biological cross-validation and state of the art machine learning not only enhances predictive performance but also enhances translational interpretability. Stratified 5-fold cross-validation was used to measure the stability of the performance, and the steady distributions of the metrics across the folds indicated that the performance would be able to tolerate moderate variability of the input data; outside cohorts might have introduced further variability that is not reflected in the current analysis. The hybrid model, ElasticNet-XGBoost is a biologically relevant, equal, and scaled, mutation prediction method that facilitates more personalized clinical treatment and the creation of biomarkers via pathways. Stratified five-fold cross-validation was used to evaluate performance stability, with the consistent distributions of metrics across folds showing good performance to relatively small variations in the input data, but external cohorts can lead to the increase in variability that is not modeled in the current analysis. The hybrid of ElasticNet and XGBoost, the hybrid ElasticNet-XGBoost is a unified, scaled, and biologically meaningful approach to mutation prediction that opens the gates to more customized clinical interventions, as well as pathway-mediated biomarker development.

While ElasticNet coefficients and SHAP values enhance transparency of model behavior, they are not to be viewed as evidence of biological or mechanistic causality, but as model-level explanations.

There are several limitations in this study. A single retrospective cohort (METABRIC) is used to perform the analysis, and the model performance is an internal cross-validation, as opposed to external validation. Generalizability to different settings may be influenced by differences in patient populations, sequencing technologies, and clinical workflows. Before wider clinical use will therefore be needed external and prospective validation. Translational This model attempts to supplement the currently available genomics testing protocols rather than replace the identification of mutations through sequencing. They can be used to aid sample selection or exploration analysis where resources are limited.

These results reveal the possible place of expression-based prediction as a complementary analysis tool and not a replacement of sequencing-based diagnosis.

5. CONCLUSION

In this study, a hybrid ElasticNetXGBoost model is proposed to predict the expression-based somatic mutation status in breast cancer, which is assessed based on the METABRIC cohort. The proposed pipeline identified TP53, PIK3CA, and MUC16 transcriptomic features reduced to compact mutation-specific gene subsets by filtering the variance, ranking mutual information, and pruning correlation. This large dimensionality reduction made it feasible to train models stably and maintain informative patterns of expression that are important to mutation status.

The hybrid model performed better than single-model baselines typical in the literature in terms of ROC-AUC (0.94, 0.92, 0.90) and precision-recall (>0.89) values. The support of model interpretability was provided using ElasticNet coefficient analysis and SHAP-based explanations on the XGBoost component, which made it possible to characterize expression features statistically correlated with mutation status. These descriptions give model-level, hypothesis-generating conclusions but not-mechanistic or causal biological conclusions.

In contrast to methods where the use of clinical variables alone or a predictive model that is not restrained by transcriptomic classifiers, the framework proposed explicitly trades off predictive performance, dimensionality control, and interpretability. The methodological rigor is further supported by the full reproducibility of the training pipeline and the experimental set up. Overall, this study shows that learning compact, mutation-related expression signatures with hybrid machine learning models is feasible and offers a complementary computational base to future research with added external validation and multi-modal genomic data.

Future efforts will be aimed at testing the proposed framework on independent cohorts, investigate ways to

combine it with other molecular and clinical data, and determine its application in various clinical settings as a computational support tool.

REFERENCES

Algamal, Z. Y., & Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Computers in Biology and Medicine, 67, 136–145. https://doi.org/10.1016/j.compbiomed.2015.10.008

Andreu-Vilarroig, C., Ceberio, J., Cortés, J.-C., de Vega, F. F., Hidalgo, J.-I., & Villanueva, R.-J. (2022). Evolutionary approach to model calibration with uncertainty. Proceedings of the Genetic and Evolutionary Computation Conference Companion, 1895–1901. https://doi.org/10.1145/3520304.3533948

Betz, M., Witz, A., Dardare, J., Michel, C., Massard, V., Boidot, R., Gilson, P., Merlin, J., & Harlé, A. (2025). Decoding mutational signatures in breast cancer: Insights from a multi-cohort study. *Translational Oncology*, *53*, 102315. https://doi.org/10.1016/j.tranon.2025.102315

Brady, S. W., Gout, A. M., & Zhang, J. (2022). Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends in Genetics, 38(2), 194–208. https://doi.org/10.1016/j.tig.2021.08.007

Breast Cancer Gene Expression Profiles (METABRIC). (n.d.). Www.Kaggle.Com.

Dinalankara, W., & Bravo, H. C. (2015). Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progression and Prognosis. Cancer Informatics, 14, CIN.S23862. https://doi.org/10.4137/CIN.S23862

Fu, X., Tan, W., Song, Q., Pei, H., & Li, J. (2022). BRCA1 and Breast Cancer: Molecular Mechanisms and Therapeutic Strategies. Frontiers in Cell and Developmental Biology, 10. https://doi.org/10.3389/fcell.2022.813457

Gale, R. P., Hochhaus, A., Cross, N. C. P., & Harrison, C. J. (2021). HGNC nomenclature for fusion genes. Leukemia, 35(11), 3039–3039. https://doi.org/10.1038/s41375-021-01437-5

Horr, C., & Buechler, S. A. (2021). Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression. Npj Breast Cancer, 7(1), 136. https://doi.org/10.1038/s41523-021-00345-2

Khan, Z., Naeem, M., Khalil, U., Khan, D. M., Aldahmani, S., & Hamraz, M. (2019). Feature Selection for Binary Classification Within Functional Genomics Experiments via Interquartile Range and Clustering. IEEE Access, 7, 78159–78169. https://doi.org/10.1109/ACCESS.2019.2922432

Kim, J. W., Lee, J., Lee, S. H., Ahn, S., & Park, K. H. (2025). Machine Learning–Based Prognostic Gene Signature for Early Triple-Negative Breast Cancer. Cancer Research and Treatment, 57(3), 731–740. https://doi.org/10.4143/crt.2024.937

Koo, N., Sharma, A. K., & Narayan, S. (2022). Therapeutics Targeting p53-MDM2 Interaction to Induce Cancer Cell Death. International Journal of Molecular Sciences, 23(9), 5005. https://doi.org/10.3390/ijms23095005

Lee, Y.-R., Chen, M., Lee, J. D., Zhang, J., Lin, S.-Y., Fu, T.-M., Chen, H., Ishikawa, T., Chiang, S.-Y., Katon, J., Zhang, Y., Shulga, Y. V., Bester, A. C., Fung, J., Monteleone, E., Wan, L., Shen, C., Hsu, C.-H., Papa, A., … Pandolfi, P. P. (2019). Reactivation of PTEN tumor suppressor for cancer treatment through inhibition of a MYC-WWP1 inhibitory pathway. Science, 364(6441). https://doi.org/10.1126/science.aau0159

Li, Q., Yang, H., Wang, P., Liu, X., Lv, K., & Ye, M. (2022). XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. Journal of Translational Medicine, 20(1), 177. https://doi.org/10.1186/s12967-022-03369-9

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777.

Mallik, S., & Zhao, Z. (2017). Towards integrated oncogenic marker recognition through mutual information - based statistically significant feature extraction: an association rule mining based study on cancer expression and methylation profiles. Quantitative Biology, 5(4), 302–327. https://doi.org/10.1007/s40484-017-0119-0

Mukherjee, A., Russell, R., Chin, S.-F., Liu, B., Rueda, O. M., Ali, H. R., Turashvili, G., Mahler-Araujo, B., Ellis, I. O., Aparicio, S., Caldas, C., & Provenzano, E. (2018). Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. Npj Breast Cancer, 4(1), 5. https://doi.org/10.1038/s41523-018-0056-8

Munkácsy, G., Santarpia, L., & Győrffy, B. (2022). Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features. Biomedicines, 10(2), 248. https://doi.org/10.3390/biomedicines10020248

Odhiambo, P., Okello, H., Wakaanya, A., Wekesa, C., & Okoth, P. (2023). Mutational signatures for breast cancer diagnosis using artificial intelligence. Journal of the Egyptian National Cancer Institute, 35(1), 14. https://doi.org/10.1186/s43046-023-00173-4

Ogundokun, R. O., Misra, S., Douglas, M., Damaševičius, R., & Maskeliūnas, R. (2022). Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks. Future Internet, 14(5), 153. https://doi.org/10.3390/fi14050153

Posta, M., & Győrffy, B. (2025). Pathway-level mutational signatures predict breast cancer outcomes and reveal therapeutic targets. *British Journal of Pharmacology*, *182*(23), 5734–5747. https://doi.org/10.1111/bph.70215

Qian, Y., Daza, J., Itzel, T., Betge, J., Zhan, T., Marmé, F., & Teufel, A. (2021). Prognostic Cancer Gene Expression Signatures: Current Status and Challenges. Cells, 10(3), 648. https://doi.org/10.3390/cells10030648

Qin, F., Luo, X., Cai, G., & Xiao, F. (2021). Shall genomic correlation structure be considered in copy number variants detection? Briefings in Bioinformatics, 22(6). https://doi.org/10.1093/bib/bbab215

Seachrist, D. D., Anstine, L. J., & Keri, R. A. (2021). FOXA1: A Pioneer of Nuclear Receptor Action in Breast Cancer. Cancers, 13(20), 5205. https://doi.org/10.3390/cancers13205205

Senbanjo, L. T., & Chellaiah, M. A. (2017). CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells. Frontiers in Cell and Developmental Biology, 5. https://doi.org/10.3389/fcell.2017.00018

Shi, H., Wu, C., Bai, T., Chen, J., Li, Y., & Wu, H. (2023). Identify essential genes based on clustering based synthetic minority oversampling technique. Computers in Biology and Medicine, 153, 106523. https://doi.org/10.1016/j.compbiomed.2022.106523

Thalor, A., Kumar Joon, H., Singh, G., Roy, S., & Gupta, D. (2022). Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. Computational and Structural Biotechnology Journal, 20, 1618–1631. https://doi.org/10.1016/j.csbj.2022.03.019

Zhang, G., Hou, S., Li, S., Wang, Y., & Cui, W. (2024). Role of STAT3 in cancer cell epithelial mesenchymal transition (Review). International Journal of Oncology, 64(5), 48. https://doi.org/10.3892/ijo.2024.5636

BIOGRAPHIES

Omji Porwal is a professor of Pharmacy in Qaiwan International University, Iraq, a top researcher in pharmaceutical sciences and computational biomedicine. His recent activity is mutation-based hybrid machine learning models to achieve precision in oncology, combining genomic signatures with AI-based predictive models of breast cancer prognosis.

Kamal Upreti is an associate professor in computer science, artificial intelligence, and bioinformatics, at Christ University, Delhi NCR. His works combine machine learning and biomedical data analytics, focusing on hybrid modeling, cancer genomics and predictive analytics in healthcare uses. His work has been highly interdisciplinary, and has had scholarly influence in fields of computational and health informatics.

Pravin R. Kshirsagar is working as a professor at J D College of Engineering & Management, Nagpur. His work is in the field of artificial intelligence, machine learning, and computational intelligence, with a specialization in biomedical signal processing, computer vision, and modeling by data. He is an author of numerous articles and books, bringing influential innovations to the field of intelligent systems and healthcare analytics.

Dr. Sarika Panwar is an Assistant Professor at the MIT Academy of Engineering, Pune, and serves as Vice Chair of the Communication Society (Pune Section). Her academic work focuses on areas such as machine learning, data analytics, and engineering education. She emphasizes the integration of modern computational techniques and model-based design approaches into teaching and research, contributing to industry-relevant innovation and skill development among engineering students.

Anurag Sharma is an Associate Professor at Newcastle University in Singapore, who works in the area of electrification, smart-grids, and the use of artificial intelligence in power systems. He has published numerous articles that have brought to light substantial contribution in sustainable energy systems and AI-powered research in power-grids.

Ganesh V. Radhakrishnan is a Professor of the Kalinga School of Management, KIIT University, Bhubaneswar. The areas he is interested in are data analytics, machine learning, and business intelligence applications. He places emphasis on the incorporation of AI-based decision models into healthcare and management systems, which adds value to the interdisciplinary innovations and sustainable analytical frameworks.

Rituraj Jain is an Assistant Professor in the Marwadi University, Rajkot, India. His studies are in the field of artificial intelligence, machine learning, and data-driven modeling, where they are applied in biomedical analytics, smart systems, and agricultural intelligence. He has authored a considerable number of publications on interdisciplinary AI.