Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 March 2026
Revised 22 April 2026
Accepted 22 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 267–280

KEYWORDS:
Basal-like breast cancer; TCGA-BRCA; machine learning; .XGBoost; SHAP; PAM50

Clinical and Genomic Summary Features for Interpretable Basal-Like Breast Cancer Identification: A TCGA-BRCA Machine Learning Study

Saad Mamoun Abdel Rahman Ahmed

Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia, smaahmed@kfu.edu.sa

ABSTRACT: Basal-like breast cancer is an aggressive molecular subtype commonly identified through PAM50 gene-expression profiling. This study examined whether low-dimensional clinical and genomic summary variables could support interpretable basal-like identification. TCGA-BRCA PanCancer Atlas data were obtained from cBioPortal. After excluding 103 patients without PAM50 annotation, 981 patients were analyzed, including 171 basal-like tumors. Logistic Regression and XGBoost were evaluated using clinical-only and clinical-genomic predictors with a stratified 80:20 split, five-fold cross-validation, class weighting, and limited hyperparameter optimization. The clinical-genomic XGBoost model achieved the strongest independent-test performance, with a ROC-AUC of 0.839, PR-AUC of 0.458, sensitivity of 82.4%, specificity of 81.6%, balanced accuracy of 81.9%, and MCC of 0.530. SHAP identified nonsynonymous tumor mutation burden, tumor break load, aneuploidy score, and MSIsensor score as the leading predictors. Results remained directionally consistent after excluding Normal-like tumors. Clinical and genomic summary variables therefore provided moderate and interpretable .basal-like identification, although external validation is required before clinical application

1. INTRODUCTION

Breast cancer is one of the most commonly diagnosed malignancies worldwide and remains a major cause of cancer-related mortality among women [1]. It is biologically heterogeneous, with molecular subtypes that differ in genomic characteristics, treatment response, and clinical outcome [2].

The PAM50 classifier categorizes breast tumors as Luminal A, Luminal B, HER2-enriched, basal-like, or Normal-like [3]. Basal-like breast cancer is characterized by aggressive clinical behavior and substantial molecular overlap with triple-negative breast cancer [4]. However, PAM50 classification depends on gene-expression measurements that may not be routinely available in settings with limited laboratory and computational resources.

Previous subtype-prediction studies have used histopathological images, RNA-sequencing data, or genome-wide copy-number profiles [5]–[7]. Although these approaches can achieve strong predictive performance, they commonly require high-dimensional inputs and complex computational processing. The clinical interpretation of their predictions may also be difficult.

This study therefore evaluated whether a small set of routinely accessible clinical characteristics and genomic summary measures could identify basal-like tumors. The main contribution is a simple comparison of clinical-only and clinical-genomic models supported by SHAP-based explanations. The study does not propose a new machine-learning algorithm or claim that the resulting model can replace PAM50 testing.

2. RELATED WORK

Intrinsic breast cancer subtypes were initially established through gene-expression profiling, demonstrating that breast tumors can be separated into biologically distinct molecular groups [2]. The PAM50 classifier subsequently provided a standardized 50-gene method for assigning intrinsic breast cancer subtypes [3]. Comprehensive TCGA analyses further confirmed substantial genomic, transcriptomic, and epigenomic heterogeneity across these subtypes [4].

Machine learning has been applied to several forms of breast cancer data for subtype prediction. Couture et al. used deep learning on histopathological images to distinguish basal-like from non-basal tumors and predict other clinicopathological characteristics [5]. Cascianelli et al. reconstructed intrinsic breast cancer subtypes using RNA-sequencing measurements and supervised machine-learning methods [6]. Xia et al. demonstrated that genome-wide copy-number alterations could predict expression-derived cancer phenotypes [7].

These approaches primarily depend on high-dimensional molecular or imaging inputs. In contrast, the present study focuses on low-dimensional clinical and genomic summary variables. It also evaluates the incremental contribution of genomic summaries beyond clinical characteristics and uses SHAP to provide interpretable model-level and patient-level explanations.

3. MATERIALS AND METHODS

3.1 Study Design and Data Source

This retrospective computational study used publicly available clinical and genomic summary data from the Breast Invasive Carcinoma cohort of The Cancer Genome Atlas PanCancer Atlas project (TCGA-BRCA). Data were obtained from cBioPortal for Cancer Genomics under the study identifier brca_tcga_pan_can_atlas_2018. cBioPortal provides harmonized, deidentified cancer-genomics datasets for reproducible secondary analysis (Cerami et al., 2012; Gao et al., 2013).

The study was designed to determine whether a limited set of demographic, histopathological, and genomic summary variables could distinguish basal-like breast cancer from other PAM50 molecular subtypes. Two predictor configurations were evaluated: a clinical model and a combined clinical-genomic model. Logistic Regression was used as an interpretable linear reference classifier, whereas XGBoost was used to capture potentially nonlinear associations and interactions.

The study workflow is presented in Figure 1.

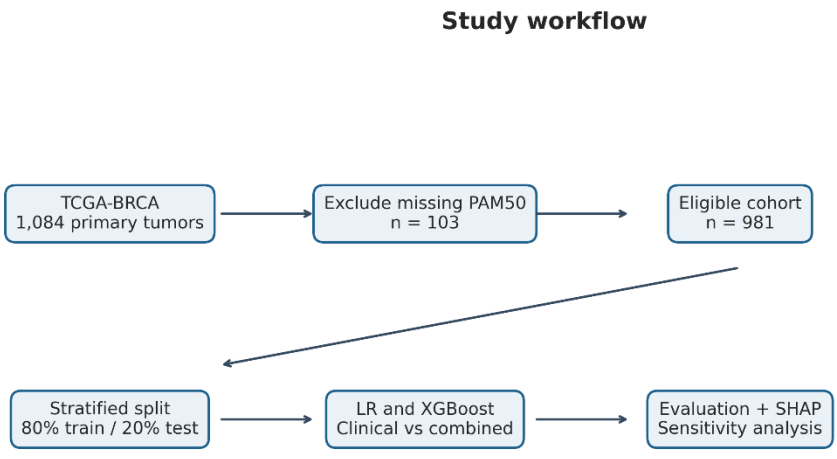

Figure 1: Study workflow. The TCGA-BRCA cohort included 1,084 primary tumors. After excluding 103 patients without PAM50 annotation, 981 patients were included and divided into stratified training and independent test cohorts.

3.2 Study Population

The downloaded TCGA-BRCA files contained 1,084 unique patients and 1,084 corresponding primary tumor samples. Patient-level and sample-level records were merged using the TCGA patient identifier (PATIENT_ID). The files were checked for duplicate identifiers before merging, and the merge was constrained to a one-to-one patient-sample relationship.

Patients were eligible when they met all the following criteria:

1. A primary breast invasive carcinoma sample was available.
2. The record represented a unique patient.
3. A PAM50 molecular subtype annotation was available.
4. The binary classification outcome could be determined.

Patients without PAM50 subtype annotation were excluded. No metastatic or recurrent samples were present in the downloaded sample file because all 1,084 records were classified as primary tumor samples. After excluding 103 patients with missing PAM50 annotations, the final analytical cohort comprised 981 patients.

The five PAM50 categories represented in the final cohort were Luminal A, Luminal B, basal-like, HER2-enriched, and Normal-like. The patient-level training and test assignments were saved to ensure that the experimental split could be reproduced exactly.

3.3 Outcome Definition

The primary outcome was binary identification of the PAM50 basal-like molecular subtype. The outcome was encoded as follows:

- Basal-like tumors: positive class, coded as 1.
- Luminal A, Luminal B, HER2-enriched, and Normal-like tumors: negative class, coded as 0.

Among the 981 eligible patients, 171 had basal-like tumors and 810 had another PAM50 subtype. Basal-like cases therefore represented 17.4% of the analytical cohort, corresponding to an approximate negative-to-positive class ratio of 4.74:1.

PAM50 subtype was used exclusively to construct the outcome and was not included as a predictor.

3.4 Predictor Variables

Two prespecified predictor configurations were investigated.

3.4.1 Clinical predictor configuration

The clinical model included:

- Age at diagnosis
- Race
- Ethnicity
- Histological tumor type

Histological tumor type was obtained from the sample-level TUMOR_TYPE variable. Sex was initially considered but was excluded after the dataset audit because all 981 PAM50-labeled patients were female. It therefore had no variance and could not contribute to classification.

3.4.2 Clinical-genomic predictor configuration

The combined model included all clinical predictors plus the following genomic summary variables:

- Aneuploidy score
- MANTIS microsatellite-instability score

- MSIsensor score
- Nonsynonymous tumor mutation burden
- Tumor break load

The genomic predictors represented summary measures rather than individual gene-expression measurements. Fraction genome altered was not included because it was unavailable in the supplied clinical files.

Variables related to pathological stage, TNM classification, treatment, survival, disease progression, recurrence, and follow-up were excluded. These variables were unnecessary for the subtype-identification objective and could introduce temporal or clinical-information leakage.

3.5 Missing-Data Assessment

Missingness was evaluated before modeling. Age, histological tumor type, MSIsensor score, nonsynonymous tumor mutation burden, and tumor break load were complete in the eligible cohort. Missing values were observed for race in 84 patients (8.6%), ethnicity in 154 patients (15.7%), aneuploidy score in 24 patients (2.4%), and MANTIS MSI score in 34 patients (3.5%).

Missing numerical values were imputed using the median calculated from the applicable training data. Missing categorical values were assigned to an explicit "Unknown" category. All imputation parameters were estimated within the training data and subsequently applied to validation or test data. The independent test set was not used to estimate imputation values.

3.6 Data Preprocessing

Numerical and categorical variables were processed through model-specific pipelines implemented with the scikit-learn Pipeline and ColumnTransformer classes.

For Logistic Regression, numerical predictors were median-imputed and standardized to a mean of zero and a standard deviation of one. Categorical predictors were imputed using the "Unknown" category and transformed using one-hot encoding. Previously unseen categories were ignored during transformation to prevent errors when processing validation or test data.

For XGBoost, numerical variables were median-imputed but not standardized because tree-based models are not sensitive to predictor scale. Categorical variables underwent the same missing-category assignment and one-hot encoding used for Logistic Regression.

Preprocessing was incorporated directly into each modeling pipeline. Consequently, imputation, standardization, and categorical encoding were fitted separately within the appropriate training folds, thereby limiting preprocessing-related data leakage.

3.7 Training and Independent Test Sets

The final cohort was divided into training and independent test sets using a fixed stratified 80:20 split with a random seed of 42. Stratification preserved the proportion of basal-like tumors in both partitions.

The resulting training set contained 784 patients, and the independent test set contained 197 patients. The test set included 34 basal-like cases. The independent test set remained unavailable to preprocessing estimation, class-weight calculation, hyperparameter selection, and model fitting.

All model development and hyperparameter optimization were conducted using the training set. The independent test set was used once for final performance evaluation.

3.8 Class-Imbalance Management

Because basal-like tumors represented a minority of the cohort, class weighting was used as the primary imbalance-management strategy. Logistic Regression was fitted using balanced class weights, which assigned weights inversely proportional to the class frequencies in the training data.

For XGBoost, the scale_pos_weight parameter was calculated as the ratio of negative to positive cases in the training partition. This increased the contribution of misclassified basal-like cases to the model-training objective.

Synthetic oversampling was not applied in the primary analysis. This avoided generating artificial observations in a relatively small dataset containing mixed numerical and categorical variables.

3.9 Machine-Learning Models

3.9.1 Logistic Regression

Logistic Regression was selected as an interpretable linear baseline. It estimates the log odds of basal-like classification as a linear combination of the processed predictors. L2 regularization was applied, and the inverse regularization-strength parameter, (C), was optimized over the values.

The maximum number of optimization iterations was set to 3,000 to support convergence after one-hot encoding.

3.9.2 XGBoost

Extreme Gradient Boosting was selected as the nonlinear classifier because it can model interactions and nonlinear relationships among clinical and genomic variables (Chen and Guestrin, 2016). The classifier used a binary logistic objective, log-loss evaluation, histogram-based tree construction, and a fixed random seed of 42.

The following hyperparameter grid was evaluated:

- Number of estimators: 100 or 250
- Maximum tree depth: 2 or 3
- Learning rate: 0.03 or 0.10
- Row subsampling proportion: 0.80
- Column subsampling proportion per tree: 0.80

The class-imbalance weight was calculated separately from the training data.

3.10 Hyperparameter Optimization and Cross-Validation

Hyperparameters were optimized using grid search with stratified five-fold cross-validation. The folds were shuffled using a random seed of 42. Average precision, equivalent to the area under the precision–recall curve for the evaluated predictions, was used as the primary tuning criterion because it is more informative than accuracy when the positive class is relatively uncommon.

For each feature configuration and classifier, the parameter combination with the highest mean cross-validated average precision was retained. Cross-validation performance was summarized using the mean and standard deviation across the five training folds for:

- ROC-AUC
- PR-AUC
- Balanced accuracy
- F1-score
- Matthews correlation coefficient
- Sensitivity

The same cross-validation structure and preprocessing principles were applied to all model configurations.

3.11 Independent Test Evaluation

After hyperparameter selection, each optimized pipeline was refitted using the complete training partition and evaluated on the untouched independent test set.

Predicted probabilities were converted into binary classifications using a fixed threshold of 0.50. The following performance measures were calculated:

- Accuracy
- Balanced accuracy

- Precision
- Sensitivity
- Specificity
- F1-score
- Matthews correlation coefficient
- ROC-AUC
- PR-AUC

Sensitivity was defined as the proportion of basal-like tumors correctly identified. Specificity was defined as the proportion of non-basal tumors correctly classified. Balanced accuracy was calculated as the arithmetic mean of sensitivity and specificity. The Matthews correlation coefficient was included because it incorporates all four elements of the confusion matrix and remains informative under class imbalance.

Because basal-like tumors represented the minority class, model assessment emphasized PR-AUC, sensitivity, balanced accuracy, F1-score, and Matthews correlation coefficient rather than accuracy alone.

3.12 Confidence Intervals

Nonparametric bootstrap resampling was used to estimate 95% confidence intervals for independent-test ROC-AUC and PR-AUC. A total of 2,000 bootstrap samples were generated by sampling test-set observations with replacement using a random seed of 42.

Bootstrap samples containing only one outcome class were discarded because ROC-AUC cannot be calculated in the absence of both positive and negative observations. Confidence limits were defined using the 2.5th and 97.5th percentiles of the valid bootstrap estimates.

3.13 Model Explainability

SHapley Additive exPlanations were used to interpret the optimized clinical-genomic XGBoost model (Lundberg and Lee, 2017). SHAP values quantify the direction and magnitude of each predictor's contribution to an individual model output relative to the model's expected prediction.

The fitted preprocessing pipeline was first applied to the independent test set. SHAP values were then calculated using TreeExplainer, which is designed for tree-based ensemble models. Explainability analysis was restricted to the clinical-genomic XGBoost model because it was the principal nonlinear model and achieved the strongest overall independent-test discrimination.

Global feature importance was calculated as the mean absolute SHAP value for each transformed predictor across all independent test observations. The explainability outputs included:

1. A SHAP summary plot showing the magnitude and direction of predictor contributions.
2. A ranked table of mean absolute SHAP values.
3. Dependence plots for the two leading transformed predictors.

SHAP results were interpreted as explanations of the fitted model rather than evidence of causal relationships.

3.14 Sensitivity Analysis

A prespecified sensitivity analysis was conducted because Normal-like PAM50 classifications can be affected by low tumor cellularity or contamination with normal breast tissue. The 36 Normal-like cases were removed, leaving 945 patients.

The binary outcome was then reconstructed as basal-like versus Luminal A, Luminal B, and HER2-enriched disease. The clinical-genomic XGBoost pipeline was retrained and retuned using the same stratified 80:20 split, five-fold cross-validation, preprocessing procedures, class weighting, hyperparameter grid, and random seed used in the primary analysis.

Performance in the sensitivity-analysis test set was assessed using the same classification and discrimination metrics. This analysis evaluated whether the principal findings were dependent on including Normal-like cases in the negative class.

3.15 Statistical Analysis of Baseline Characteristics

Baseline characteristics were summarized for the complete cohort and separately for basal-like and non-basal tumors. Continuous variables were summarized using medians and interquartile ranges because several genomic measures exhibited skewed distributions. Categorical variables were reported as counts and percentages based on available observations.

Continuous variables were compared using the two-sided Mann–Whitney U test. Categorical distributions were evaluated using the chi-square test of independence. The statistical comparisons were exploratory and were not used for feature selection. A two-sided ($P<0.05$) was considered statistically significant. No adjustment for multiple comparisons was applied; therefore, baseline (P)-values were interpreted descriptively.

3.16 Software and Reproducibility

The analysis was conducted in Python 3.12. Data processing and statistical analysis used pandas, NumPy, and SciPy. Machine-learning pipelines, preprocessing, cross-validation, and performance evaluation were implemented using scikit-learn. XGBoost was used for gradient-boosted classification, and SHAP was used for model interpretation. Matplotlib and Seaborn were used to generate publication-quality figures.

Reproducibility was supported by:

- A fixed random seed of 42
- A saved patient-level training and test assignment
- Pipeline-based preprocessing
- Explicit hyperparameter grids
- A machine-readable analysis summary
- Reproducible Python source code
- Exported primary and supplementary result tables

3.17 Ethical Considerations

The study used publicly available, deidentified data from TCGA and cBioPortal. No direct contact with participants occurred, and no personally identifiable information was accessed. The analysis therefore constituted secondary analysis of publicly accessible anonymized data and did not require additional institutional review board approval or individual informed consent. The original TCGA data collection was conducted under the participating institutions' ethical-review and informed-consent procedures.

4. RESULTS

4.1 Cohort Characteristics

The original dataset contained 1,084 primary breast cancer cases. After excluding 103 patients without PAM50 annotation, 981 patients remained. The cohort included 499 Luminal A, 197 Luminal B, 171 basal-like, 78 HER2-enriched, and 36 Normal-like tumors (Fig. 2). Basal-like tumors represented 17.4% of the final cohort.

Patients with basal-like tumors were younger than those with other subtypes (median: 54.0 versus 59.0 years, ($P=0.001$)). They also had higher aneuploidy scores, TMB, MSIsensor scores, MANTIS MSI scores, and tumor break loads (all ($P<0.001$)). Infiltrating ductal carcinoma was more frequent among basal-like tumors (86.5% versus 71.0%, ($P<0.001$)). Baseline characteristics are summarized in Table I.

Table I: Baseline characteristics of the TCGA-BRCA cohort according to basal-like subtype status.

Characteristic	Overall (n=981)	Basal-like (n=171)	Other (n=810)	P value
Age, years	58.00 (48.00â€“67.00)	54.00 (47.50â€“63.50)	59.00 (49.00â€–68.00)	0.001
Aneuploidy score	11.00 (5.00â€–19.00)	15.00 (10.50â€–21.00)	10.00 (5.00â€–18.00)	<0.001
MANTIS MSI	0.30 (0.28â€–0.32)	0.31 (0.29â€–0.33)	0.29 (0.27â€–0.32)	<0.001
MSIsensor	0.20 (0.04â€–0.54)	0.53 (0.17â€–1.21)	0.17 (0.03â€–0.42)	<0.001
TMB, nonsynonymous	1.40 (0.90â€–2.40)	2.37 (1.55â€–3.67)	1.23 (0.83â€–2.03)	<0.001
Tumor break load	83.00 (27.00â€–193.00)	170.00 (98.50â€–271.00)	65.00 (21.25â€–165.00)	<0.001
White race	678 (75.6%)	106 (63.9%)	572 (78.2%)	<0.001
Not Hispanic/Latino	796 (96.3%)	149 (98.0%)	647 (95.9%)	0.083
Infiltrating ductal carcinoma	723 (73.7%)	148 (86.5%)	575 (71.0%)	<0.001

Continuous variables are presented as median (interquartile range), and categorical variables as number (percentage). P values were calculated using the Mann–Whitney U test or chi-square test, as appropriate.

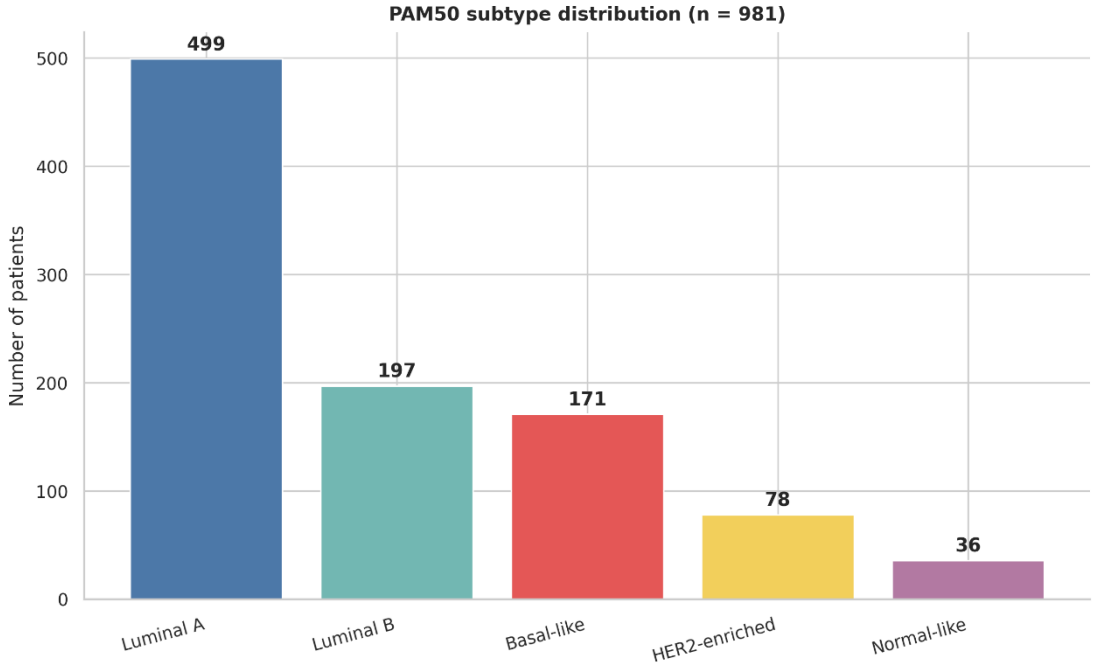

Figure 2: Distribution of PAM50 molecular subtypes in the eligible TCGA-BRCA cohort.

The final cohort comprised 499 Luminal A, 197 Luminal B, 171 basal-like, 78 HER2-enriched, and 36 Normal-like tumors.

4.2 Data Completeness

Age, histological tumor type, TMB, MSIsensor score, and tumor break load were complete. Missingness was limited for aneuploidy score (2.4%) and MANTIS MSI score (3.5%) but was higher for race (8.6%) and ethnicity (15.7%). Sex was excluded because all eligible patients were female. Predictor availability is presented in Table II.

Table I:. Predictor definitions and missing-data distribution in the eligible cohort.

Variable	Dataset column	Type	Missing, n	Missing, %
Age at diagnosis	AGE	Numerical	0	0
Race	RACE	Categorical	84	8.562691
Ethnicity	ETHNICITY	Categorical	154	15.69827
Histological tumor type	TUMOR_TYPE	Categorical	0	0
Aneuploidy score	ANEUPLOIDY_SCORE	Numerical	24	2.446483
MANTIS MSI score	MSI_SCORE_MANTIS	Numerical	34	3.465851
MSIsensor score	MSI_SENSOR_SCORE	Numerical	0	0
Tumor mutation burden (nonsynonymous)	TMB_NONSYNONYMOUS	Numerical	0	0
Tumor break load	TBL_SCORE	Numerical	0	0

Numerical missing values were median-imputed, while missing categorical values were assigned to an Unknown category using training data only.

4.3 Cross-Validation Performance

The clinical-genomic XGBoost model achieved the strongest cross-validation performance, with a mean ROC-AUC of 0.849 ± 0.070 and PR-AUC of 0.601 ± 0.130. Its mean balanced accuracy was 0.752, compared with 0.597 for clinical-only XGBoost.

Adding genomic variables also improved Logistic Regression. Its mean ROC-AUC increased from 0.691 to 0.764, while PR-AUC increased from 0.370 to 0.484. Complete cross-validation results and selected hyperparameters are provided in Table III.

Table III: Five-fold cross-validation performance of Logistic Regression and XGBoost using clinical and clinical-genomic predictor configurations.

Feature set	Model	Best parameters	ROC AUC mean	ROC AUC SD	PR AUC mean	PR AUC SD	BALANCED ACCURACY mean	BALANCED ACCURACY SD	F1 mean	F1 SD	MCC mean	MCC SD	SENSITIVITY mean	SENSITIVITY SD
Clinical	Logistic Regression	{"model__C": 10.0}	0.6908	0.082728	0.369841	0.117	0.620046	0.070642	0.359828	0.062127	0.182913	0.10768	0.706878	0.143168
Clinical	XGBoost	{"model__colsample_bytree": 0.8, "model__learning_rate": 0.03, "model__max_depth": 2, "model__n_estimators": 100, "model__subsample": 0.8}	0.666684	0.072215	0.322326	0.09562	0.596857	0.061617	0.342855	0.05459	0.147954	0.094253	0.641799	0.090366
Clinical + genomic	Logistic Regression	{"model__C": 10.0}	0.763799	0.092081	0.48449	0.148933	0.695089	0.086148	0.44571	0.091602	0.307685	0.137016	0.699206	0.124133

Clinical + genomic	XGBoost	{"model__colsample_bytree": 0.8, "model__learning_rate": 0.03, "model__max_depth": 2, "model__n_estimators": 250, "model__subsample": 0.8}	0.849016	0.069942	0.600992	0.129756	0.751603	0.085463	0.526781	0.113283	0.41697	0.151073	0.728836	0.126556

Values represent mean ± standard deviation across the five training folds.

4.4 Independent Test Performance

The independent test set contained 197 patients, including 34 basal-like tumors. The clinical-genomic XGBoost model achieved the strongest overall performance, with an ROC-AUC of 0.839 (95% CI: 0.765–0.905) and PR-AUC of 0.458 (95% CI: 0.329–0.637). At the predefined 0.50 threshold, it achieved 82.4% sensitivity, 81.6% specificity, 81.9% balanced accuracy, an F1-score of 0.609, and an MCC of 0.530.

The model correctly identified 28 of the 34 basal-like tumors, with six false-negative and 30 false-positive classifications. Clinical-only Logistic Regression produced the highest sensitivity at 88.2%, but its lower specificity of 58.9% resulted in weaker overall discrimination.

Adding genomic summary variables improved both classifiers. For XGBoost, ROC-AUC increased from 0.758 to 0.839 and PR-AUC increased from 0.334 to 0.458. Independent test performance is summarized in Table IV, while the ROC and precision–recall curves are shown in Fig. 3.

Table IV: Performance of the optimized models in the independent test cohort.

Feature set	Model	Accuracy	Balanced Accuracy	Precision	Sensitivity	Specificity	F1	MCC	ROC-AUC	ROC-AUC 95% CI	PR-AUC	PR-AUC 95% CI	TN	FP	FN	TP
Clinical	Logistic Regression	0.639594	0.735655	0.309278	0.882353	0.588957	0.458015	0.356249	0.763984	0.682 â€“0.837	0.372295	0.244 â€“0.524	96	67	4	30
Clinical	XGBoost	0.639594	0.712378	0.301075	0.823529	0.601227	0.440945	0.321525	0.757939	0.677 â€“0.828	0.334147	0.227 â€“0.482	98	65	6	28
Clinical + genomic	Logistic Regression	0.730964	0.755955	0.369863	0.794118	0.717791	0.504673	0.400547	0.801335	0.728 â€“0.869	0.376478	0.264 â€“0.552	117	46	7	27
Clinical + genomic	XGBoost	0.817259	0.81974	0.482759	0.823529	0.815951	0.608696	0.530199	0.838867	0.765 â€“0.905	0.458125	0.329 â€“0.637	133	30	6	28

The test cohort contained 197 patients, including 34 basal-like tumors. Confidence intervals were estimated using 2,000 bootstrap samples.

Natural Resources for Human Health
eISSN: 2583-1194

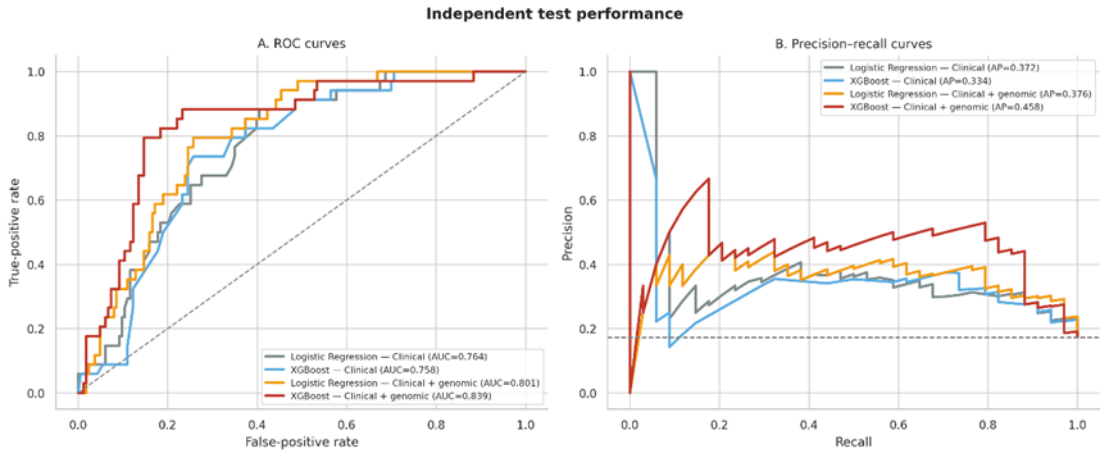

Figure 3: Independent test performance of the evaluated models. (A) Receiver operating characteristic curves. (B) Precision–recall curves. The clinical-genomic XGBoost model achieved the highest ROC-AUC and PR-AUC.

4.5 SHAP Interpretation

SHAP analysis identified nonsynonymous TMB as the most influential predictor in the clinical-genomic XGBoost model, followed by tumor break load, aneuploidy score, and MSIsensor score. Higher values of these genomic variables generally increased the model's predicted probability of basal-like disease.

Infiltrating lobular carcinoma was associated with lower basal-like probability, whereas infiltrating ductal carcinoma predominated among basal-like cases. Age and MANTIS MSI score made smaller contributions. The global SHAP distribution is presented in Fig. 4, and dependence patterns for TMB and tumor break load are shown in Fig. 5.

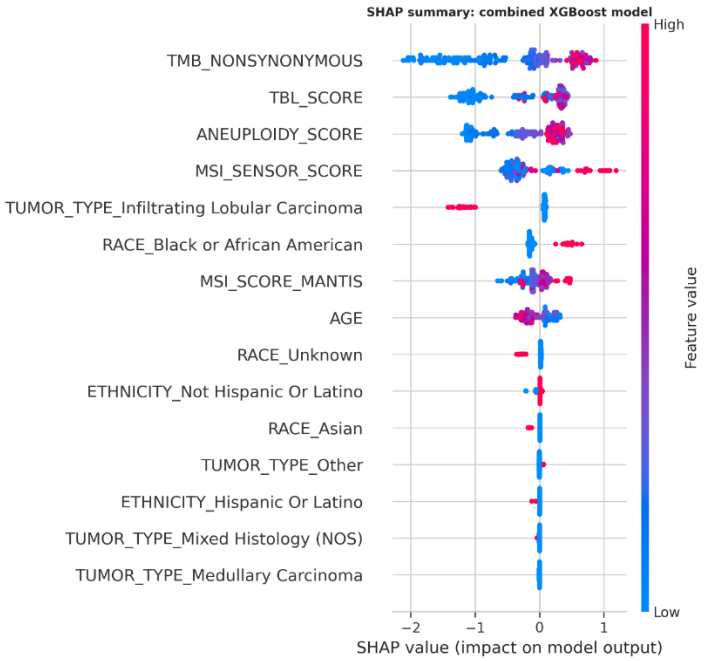

Figure 4: SHAP summary plot for the clinical-genomic XGBoost model. Predictors are ordered by mean absolute SHAP value. Positive SHAP values increase the predicted probability of basal-like disease, whereas negative values reduce it. Color represents the relative predictor value.

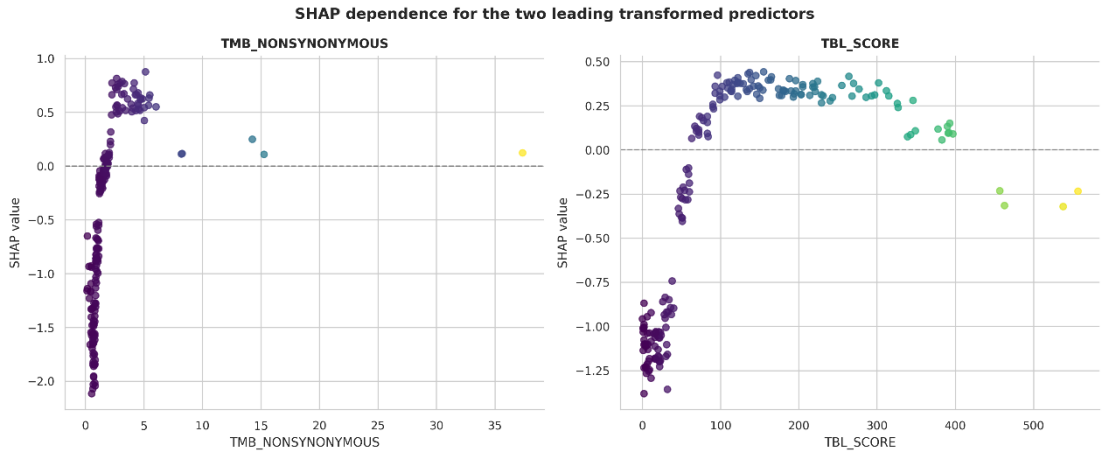

Figure 5: SHAP dependence plots for the two leading predictors. The plots show the relationships between nonsynonymous tumor mutation burden, tumor break load, and their contributions to basal-like classification.

4.6 Sensitivity Analysis

After excluding 36 Normal-like tumors, 945 patients remained. The retrained clinical-genomic XGBoost model achieved an ROC-AUC of 0.803 and PR-AUC of 0.484. Its sensitivity was 67.6%, specificity was 80.0%, and balanced accuracy was 73.8%.

Although sensitivity decreased, the model retained moderate discrimination, indicating that its main predictive pattern was not dependent exclusively on including Normal-like tumors in the negative class. The complete sensitivity-analysis results are provided in Supplementary Table S1.

Supplementary Table S1: Sensitivity analysis after excluding Normal-like tumors.

Accuracy	Balanced Accuracy	Precision	Sensitivity	Specificity	F1	MCC	ROC-AUC	PR-AUC	TN	FP	FN	TP	Cohort	N	Test N	Basal test N	Best parameters
0.777778	0.738235	0.425926	0.676471	0.8	0.522727	0.405114	0.803416	0.48379	124	31	11	23	Normal-like excluded	945	189	34	{"model__colsample_bytree": 0.8, "model__learning_rate": 0.03, "model__max_depth": 3, "model__n_estimators": 250, "model__subsample": 0.8}

The clinical-genomic XGBoost model was retrained and evaluated using the same analytical procedure applied in the primary analysis.

5. DISCUSSION

The principal finding was that adding five genomic summary measures improved the observed predictive performance of both Logistic Regression and XGBoost. The clinical-genomic XGBoost model achieved the strongest performance, balancing sensitivity and specificity while producing a PR-AUC substantially higher than the basal-like prevalence of 17.4%. These findings suggest that genomic summary variables contain incremental predictive information beyond demographic and histological characteristics.

The SHAP ranking was biologically plausible. Basal-like tumors exhibit substantial genomic instability and distinctive mutation profiles [4]. Previous research has also reported comparatively elevated tumor mutational burden in basal-like breast cancer [12]. Tumor break load and aneuploidy score summarize structural and chromosomal disruption, while tumor break load has been proposed as a biologically relevant measure of genomic instability [13]. These results suggest that the model identified genomic characteristics consistent with established basal-like tumor biology.

Race also contributed to some model predictions. However, this association must not be interpreted causally. It may reflect population structure, differences in the TCGA sampling process, unmeasured socioeconomic or biological factors, or correlations with other predictors. External validation across more diverse cohorts is necessary before drawing conclusions about demographic contributions.

Unlike previous image-based, RNA-sequencing, or genome-wide models [5]–[7], the present study used only nine predictors. Its main contribution is therefore simplicity and interpretability rather than maximum predictive accuracy. Such low-dimensional models may provide useful screening or research tools when comprehensive gene-expression profiling is unavailable. Nevertheless, the proposed model is not a substitute for PAM50 because PAM50 defined the reference outcome and no external clinical validation was performed.

5.1 Limitations

This study has several limitations. First, it was based on a retrospective single-cohort analysis and included only 34 basal-like patients in the independent test set. Second, race and ethnicity contained missing values, while all PAM50-labeled patients were female. Third, predictions were evaluated using a fixed probability threshold of 0.50 without formal threshold optimization or calibration assessment. Fourth, no independent external cohort was available. Finally, SHAP explains the behavior of the fitted model but does not establish biological causality. Although the sensitivity analysis supported the main findings, model sensitivity decreased after Normal-like tumors were excluded.

6. CONCLUSION

A small set of clinical and genomic summary variables provided moderate and interpretable identification of basal-like breast cancer in the TCGA-BRCA cohort. The clinical-genomic XGBoost model produced the strongest observed performance, while SHAP highlighted tumor mutation burden, tumor break load, aneuploidy score, and MSIsensor score as the leading predictors. These findings support the predictive value of genomic summary measurements while maintaining a relatively simple and explainable modeling framework. Independent and prospective validation is required before clinical application.

DECLARATIONS

Ethics approval and consent to participate: Publicly available, deidentified data were used. Additional institutional ethical approval and participant consent were therefore not required.

Data availability: The data analyzed in this study are publicly available through cBioPortal under the TCGA-BRCA PanCancer Atlas study identifier brca_tcga_pan_can_atlas_2018.

Code availability: The complete analysis code, model configurations, and patient-level split manifests are available with the study materials.

Competing interests: The authors declare that they have no competing interests.

Acknowledgments: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. **KFU264311**]

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] C. M. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, 2000, doi: 10.1038/35021093.

[3] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, 2009, doi: 10.1200/JCO.2008.18.1370.

[4] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012, doi: 10.1038/nature11412.

[5] H. D. Couture *et al.*, "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype," *npj Breast Cancer*, vol. 4, Art. no. 30, 2018, doi: 10.1038/s41523-018-0079-1.

[6] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, and E. Medico, "Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer," *Sci. Rep.*, vol. 10, Art. no. 14071, 2020, doi: 10.1038/s41598-020-70832-2.

[7] Y. Xia, C. Fan, K. A. Hoadley, J. S. Parker, and C. M. Perou, "Genetic determinants of the molecular portraits of epithelial cancers," *Nat. Commun.*, vol. 10, Art. no. 5666, 2019, doi: 10.1038/s41467-019-13588-2.

[8] E. Cerami *et al.*, "The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discov.*, vol. 2, no. 5, pp. 401–404, 2012, doi: 10.1158/2159-8290.CD-12-0095.

[9] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, Art. no. pl1, 2013, doi: 10.1126/scisignal.2004088.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[12] A. Thomas *et al.*, "Tumor mutational burden is a determinant of immune-mediated survival in breast cancer," *Oncoimmunology*, vol. 7, no. 10, Art. no. e1490854, 2018, doi: 10.1080/2162402X.2018.1490854.

[13] S. Lakbir *et al.*, "Tumour break load is a biologically relevant feature of genomic instability with prognostic value in colorectal cancer," *Eur. J. Cancer*, vol. 177, pp. 94–102, 2022, doi: 10.1016/j.ejca.2022.09.034.