Annals of Mathematics and Computer Science ISSN: 2789-7206
Vol 36 (2026) 1-24 https://doi.org/10.56947/amcs.v36.911

LEAKAGE-AWARE EXPLAINABLE FRAMEWORK FOR BREAST CANCER PROGNOSIS

TOMILOLA OLADOYIN AJOSE[1*] CHINWE PEACE IGIRI[2*], OJEN KUMAR NARAIN[3], AND ADHIR MAHARAJ[4]

Abstract. Breast cancer remains a leading cause of cancer-related mortality among women, necessitating reliable and interpretable prognostic models for personalized treatment planning. This study developed a leakage-aware, explainable machine learning framework for breast cancer survival prediction using clinicopathological data from 4,024 patients in the SEER database. The framework integrated feature engineering, leakage detection, hyperparameter optimization, calibration assessment, and explainable artificial intelligence. Among five evaluated algorithms, CatBoost achieved the best performance, with a ROC-AUC of 0.725 and a five-fold cross-validated ROC-AUC of 0.746 (95% CI: 0.708–0.785). SHAP analysis identified Node Ratio, Age, Hormone Index, Tumor Burden, and Grade as the most influential predictors. The proposed framework provides transparent, reliable, and clinically relevant prognostic predictions for breast cancer risk stratification.

Keywords. Breast cancer; Survival prediction; Machine learning; Explainable artificial intelligence; CAT Boost; SHAP; Feature engineering.
2020 Mathematics Subject Classification. Primary 46L55; Secondary 44B20

1. Introduction and preliminaries

Breast cancer remains a major public health challenge worldwide. According to global estimates, it was the most frequently diagnosed cancer, with approximately 2.3 million new cases and nearly 685,000 deaths reported in 2020 [17]. Despite substantial advances in screening, early diagnosis, and treatment, breast cancer continues to be the most common malignancy among women and one of the leading causes of cancer-related mortality globally [16]. The disease burden remains particularly high in low- and middle-income countries, underscoring the need for reliable prognostic tools that can support personalized treatment planning and improve patient outcomes.

Date: Received: Jul 10, 2026; Revised: Aug 9, 2026; Accepted: Aug 14, 2026.
[1] Corresponding author

Survival prediction plays a central role in breast cancer management by informing treatment decisions, risk stratification, follow-up planning, and shared decision-making between clinicians and patients [5]. Conventional prognostic approaches, including the Tumor–Node–Metastasis (TNM) staging system and Cox proportional hazards models, have demonstrated clinical utility but are limited in their ability to capture complex nonlinear relationships and interactions among demographic, pathological, and molecular characteristics [5]. Breast cancer prognosis is influenced by multiple interrelated factors, including tumor type, nodal involvement, hormone receptor status, patient age, and treatment response [9].

Recent advances in artificial intelligence (AI) and machine learning (ML) have provided powerful computational approaches for modelling complex clinical relationships. Machine learning algorithms, including Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, have demonstrated promising performance in breast cancer recurrence prediction, treatment outcome assessment, and survival prediction [6, 9, 14]. These algorithms can automatically identify complex patterns within high-dimensional clinical data and frequently outperform conventional statistical models in predictive tasks [13].

Despite these advances, several methodological challenges continue to limit the development of clinically reliable prognostic models. First, many published studies inadvertently include predictor variables containing information that would not be available at the time of prediction, resulting in information leakage and artificially inflated model performance [8]. Information leakage occurs when predictor variables contain information about the target outcome that would be unavailable during real-world model deployment, thereby compromising model validity and generalizability [15, 17]. Second, numerous studies focus primarily on discrimination metrics, such as accuracy and the area under the receiver operating characteristic curve (ROC–AUC), while overlooking calibration assessment, statistical comparison of competing models, and external validation [18]. Third, the increasing complexity of modern ML algorithms has raised concerns regarding model transparency and interpretability, thereby limiting clinicians' confidence in adopting these systems for routine clinical practice [2].

The emergence of explainable artificial intelligence (XAI) has provided practical tools for addressing these limitations. Among these, Shapley Additive Explanations (SHAP) quantify the contribution of each predictor to individual model predictions, thereby improving transparency and facilitating clinically meaningful interpretation of machine learning outputs. Such explainability is essential for promoting clinician trust and supporting evidence-based decision-making.

Another important yet underexplored aspect of breast cancer prognostic modelling is clinically guided feature engineering. Existing clinicopathological variables can be transformed into composite biomarkers that better represent disease burden than individual predictors alone. For example, the lymph node ratio has

consistently demonstrated superior prognostic value compared with conventional nodal staging in several breast cancer populations. Similarly, composite measures representing tumor burden and hormone receptor activity may provide additional prognostic information beyond routinely recorded clinical variables.

Furthermore, recent reporting guidelines, including the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Using Artificial Intelligence (TRIPOD+AI) Statement, emphasize the importance of explainability, rigorous validation, calibration assessment, and transparent reporting when developing machine learning-based clinical prediction models [2, 10, 11]. Adherence to these recommendations is increasingly recognized as essential for producing reproducible, trustworthy, and clinically applicable prognostic models.

Therefore, this study proposes a clinically guided, leakage-aware, explainable machine learning framework for breast cancer survival prediction using routinely collected clinicopathological data from the Surveillance, Epidemiology, and End Results (SEER) database. The proposed framework integrates data quality assessment, information leakage detection, clinically guided feature engineering, hyperparameter optimization, calibration assessment, statistical model comparison, and explainability analysis within a reproducible modelling pipeline. Four clinically engineered prognostic variables—*Node Ratio*, *Tumor Burden*, *Hormone Index*, and *Node Density*—were developed and incorporated with conventional clinicopathological variables to improve predictive performance and clinical interpretability. Multiple machine learning algorithms were systematically evaluated using stratified cross-validation and independent test-set evaluation. Finally, SHAP-based explainability analysis was performed to identify the most influential prognostic predictors and provide clinically meaningful insights into model behaviour.

2. Materials and Methods

2.1. **Study Design and Analytical Framework.** This study proposes a clinically guided, leakage-aware, and explainable machine learning framework for breast cancer survival prediction using routinely collected clinicopathological data. The framework was designed to address major methodological limitations commonly reported in breast cancer prognostic modelling, including information leakage, limited model interpretability, inadequate calibration assessment, and insufficient statistical validation. Rather than focusing solely on predictive performance, the proposed framework integrates clinically guided feature engineering, rigorous model development, explainable artificial intelligence (XAI), calibration analysis, and statistical comparison to improve the reliability and transparency of prognostic prediction.

The analytical workflow consisted of twelve sequential phases, namely: (i) data acquisition, (ii) exploratory data analysis, (iii) data quality assessment, (iv) information leakage assessment, (v) feature engineering, (vi) data preparation, (vii)

machine learning model development, (viii) hyperparameter optimization, (ix) model evaluation, (x) statistical comparison, (xi) explainability analysis, and (xii) calibration assessment. Each stage was designed to ensure methodological rigor, reproducibility, and clinical interpretability.

The entire analytical pipeline was implemented in Python 3.12 using Google Colaboratory. Open-source libraries including Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, CatBoost, SHAP, Matplotlib, and Seaborn were employed for data preprocessing, model development, evaluation, visualization, and explainability analysis. Figure 1 illustrates the overall analytical framework adopted in this study.

2.2. **Phase I: Dataset Acquisition and Description.** The study utilized the publicly available Surveillance, Epidemiology, and End Results (SEER) Breast Cancer database maintained by the United States National Cancer Institute. The SEER program is one of the largest population-based cancer registries worldwide and is widely used for prognostic modelling because of its standardized data collection procedures, comprehensive clinicopathological information, and high-quality follow-up records [3, 16, 15]. After data preprocessing, a total of 4,024 breast cancer patients were included in the analysis. Overall survival served as the study outcome and was formulated as a binary classification problem, where patients alive at the end of follow-up were assigned a value of 0 and deceased patients a value of 1. The predictor variables comprised routinely collected demographic and clinicopathological characteristics, including age, race, marital status, tumor stage, nodal stage, histological grade, tumor size, hormone receptor status, number of examined lymph nodes, and number of positive lymph nodes.

2.3. **Phase II: Exploratory Data Analysis.** Exploratory data analysis (EDA) was performed to understand the characteristics of the dataset before model development. Descriptive statistics were computed for all study variables. Frequencies and percentages summarized categorical variables, whereas appropriate measures of central tendency and dispersion were calculated for continuous variables. The distribution of the outcome variable was examined to assess class imbalance. Of the 4,024 patients, 3,408 (84.69%) survived and 616 (15.31%) died during follow-up, indicating substantial class imbalance. Pearson's correlation coefficient was computed to evaluate relationships among predictor variables and identify potential multicollinearity. Strong positive correlations were observed between N Stage and Sixth AJCC Stage ($r = 0.882$), and between N Stage and the number of positive lymph nodes ($r = 0.838$). These relationships informed subsequent feature engineering and interpretation of model outputs. Data visualization, including class distribution plots and correlation heatmaps, was also employed to support exploratory analysis and modelling decisions.

2.4. **Phase III: Data Quality Assessment.** A comprehensive data quality assessment was conducted before model development. All variables were examined for missing values, duplicate records, inconsistencies, and implausible observations. No missing values or duplicate records were identified; therefore, no data

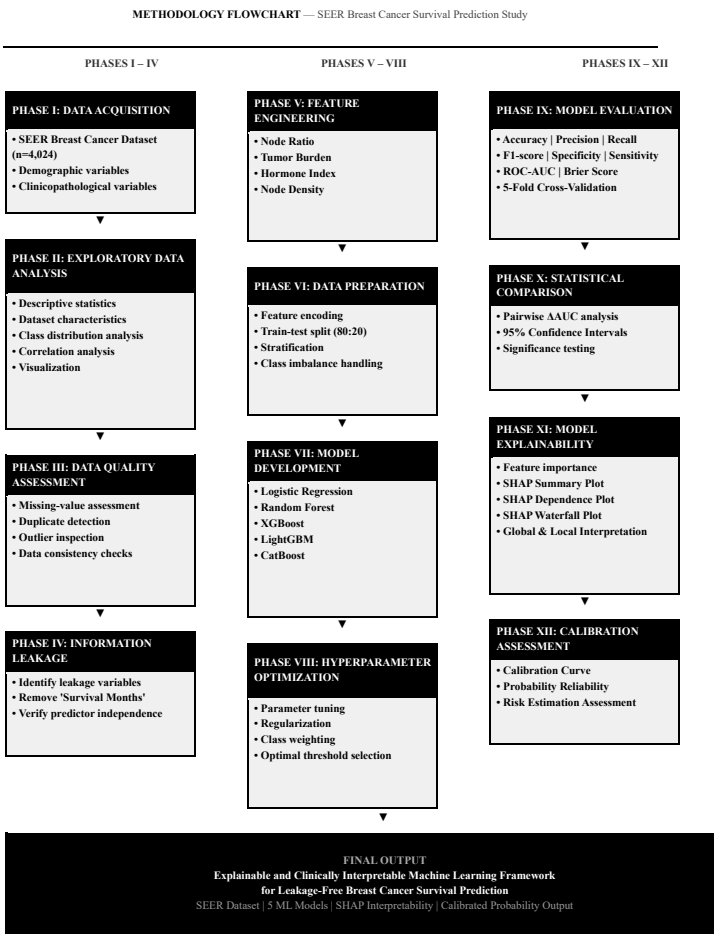

FIGURE 1. Overview of the proposed clinically guided leakage-aware explainable machine learning framework for breast cancer survival prediction. The framework comprises twelve sequential phases, including data acquisition, exploratory data analysis, data quality assessment, information leakage assessment, clinically guided feature engineering, data preparation, machine learning model development, hyperparameter optimization, model evaluation, statistical comparison, explainability analysis, and calibration assessment.

imputation or duplicate removal was required. Numerical variables were further inspected to verify plausible ranges and consistency. The absence of major data quality issues confirmed the suitability of the dataset for predictive modelling without extensive preprocessing.

2.5. **Phase IV: Information Leakage Assessment.** Information leakage is a major source of overly optimistic performance estimates in predictive modelling because it allows models to exploit information unavailable during real-world prediction [8, 6, 7]. To mitigate this risk, all predictor variables were systematically evaluated before model development. The variable *Survival Months* was identified

as a leakage variable because it directly represents patient survival time and would not be available when making prognostic predictions at diagnosis. Including this variable would artificially inflate predictive performance and compromise model generalizability. Consequently, *Survival Months* was excluded from all modelling procedures. This leakage-aware strategy ensured that the developed models more accurately reflected real-world clinical prediction scenarios.

2.6. **Phase V: Feature Engineering.** Clinically guided feature engineering was performed to derive composite prognostic variables from routinely available clinicopathological data. Four engineered features were developed. *Node Ratio* was calculated as the ratio of positive lymph nodes to the total number of examined lymph nodes, providing a standardized measure of nodal involvement. *Tumor Burden* was computed by multiplying tumor size by the number of positive lymph nodes to capture both tumor extent and metastatic spread. *Hormone Index* combined estrogen receptor and progesterone receptor status into a single variable representing overall hormone receptor positivity. Finally, *Node Density* was calculated as the ratio of positive lymph nodes to tumor size, reflecting metastatic burden relative to tumor volume. These engineered variables demonstrated substantial variability across patients and provided additional prognostic information beyond conventional clinicopathological variables.

2.7. **Phase VI: Data Partitioning and Class Imbalance Management.** The dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the original class distribution. Because mortality cases constituted only 15.31% of the study population, class imbalance was addressed during model training using class weighting. For the CatBoost classifier, class weights were automatically assigned according to class frequencies to improve minority-class recognition while minimizing bias toward the majority class.

2.8. **Phase VII: Machine Learning Model Development.** Five supervised machine learning algorithms were developed and compared: Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost. These algorithms were selected to represent both conventional statistical learning and modern ensemble learning approaches that have demonstrated strong performance in clinical prediction studies. All models were trained using identical training and testing partitions to ensure a fair comparative evaluation.

2.9. **Phase VIII: Hyperparameter Optimization.** Hyperparameter optimization was performed to maximize predictive performance while reducing overfitting. For the CatBoost classifier, Bayesian optimization identified the optimal hyperparameter configuration consisting of 875 boosting iterations, tree depth of 4, learning rate of 0.0103, and an L2 regularization coefficient of 16.726. These parameters provided the best balance between model complexity and predictive accuracy.

2.10. **Phase IX: Model Evaluation.** Model performance was evaluated on the independent test set using multiple discrimination and calibration metrics. Discrimination was assessed using accuracy, precision, recall, specificity, F1-score, and the area under the receiver operating characteristic curve (ROC–AUC). Calibration performance was evaluated using the Brier score and calibration curves to assess agreement between predicted probabilities and observed outcomes. The optimized CatBoost model achieved an accuracy of 0.773, recall of 0.553, F1-score of 0.426, and ROC–AUC of 0.725. Model robustness was further assessed using five-fold stratified cross-validation, which yielded a mean ROC–AUC of 0.746 (95% CI: 0.708–0.785), indicating stable predictive performance across validation folds.

2.11. **Phase X: Statistical Comparison of Models.** Statistical comparison of model performance was conducted using pairwise comparisons of cross-validated ROC–AUC values. The CatBoost classifier demonstrated statistically significant improvements over Random Forest, XGBoost, and LightGBM, whereas no statistically significant difference was observed between CatBoost and Logistic Regression. This analysis provided additional evidence supporting the robustness of the selected model.

2.12. **Phase XI: Explainability Analysis.** Model explainability was investigated using SHapley Additive exPlanations (SHAP), which provide both global and local interpretations of machine learning predictions. SHAP values quantified the contribution of each predictor to the predicted survival outcome. Global feature importance analysis identified *Node Ratio*, *Age*, *Hormone Index*, *Tumor Burden*, and *Histological Grade* as the most influential predictors. SHAP summary plots, dependence plots, and waterfall plots were generated to visualize feature effects and facilitate clinical interpretation of model behaviour.

2.13. **Phase XII: Calibration Assessment.** Calibration analysis was performed to evaluate the reliability of predicted probabilities. Calibration curves and the Brier score were used to assess agreement between predicted mortality risk and observed outcomes. Accurate calibration is essential for clinical decision support because it ensures that predicted probabilities closely reflect true patient risk, thereby improving risk stratification and supporting individualized treatment planning.

3. Methodology Framework and Interpretability Architecture

To bridge the critical gap between high-capacity predictive modeling and rigorous interpretability, our theoretical framework integrates a primary gradient-boosted decision tree ensemble (CatBoost) with an inherently interpretable, basis-expanded Generalized Additive Model (GAM). While post-hoc attribution methods such as SHAP (SHapley Additive exPlanations) provide valuable local instance-level heuristics, they lack global structural guarantees and exact mathematical fidelity bounds.

In this section, we formally define our engineered feature space, construct an inherently interpretable GAM architecture over this domain, and establish a rigorous mathematical framework bounding the approximation error between the black-box ensemble and the interpretable surrogate.

3.1. **Engineered Feature Domain and Basis Function Construction.** Let $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T \in \mathcal{X} \subset \mathbb{R}^d$ represent a feature vector defined over a compact hyperrectangular domain $\mathcal{X} = \prod_{i=1}^{d}[a_i, b_i]$, where $a_i, b_i \in \mathbb{R}$ with $a_i < b_i$. To capture complex non-linear functional forms without sacrificing structural transparency, we construct a normalized K-dimensional B-spline basis (or orthogonal Chebyshev polynomial basis) for each feature dimension $i \in \{1, \ldots, d\}$. Let $\mathbf{\Phi}_i(x_i) = [\phi_{i,1}(x_i), \phi_{i,2}(x_i), \ldots, \phi_{i,K}(x_i)]^T$ denote the vector of basis functions defined over a uniform grid mesh with knot spacing:

$$h = \max_{1 \le i \le d} \max_{1 \le k \le K-1} |x_{i,k+1} - x_{i,k}|. \tag{3.1}$$

For paired feature interactions identified through domain-specific feature engineering, let $\mathcal{I} \subset \{(i, j) \mid 1 \le i < j \le d\}$ denote the set of active interaction indices. The bivariate tensor-product basis for any pair $(i, j) \in \mathcal{I}$ is given by:

$$\mathbf{\Psi}_{ij}(x_i, x_j) = \mathbf{\Phi}_i(x_i) \otimes \mathbf{\Phi}_j(x_j) = \{\phi_{i,k}(x_i) \cdot \phi_{j,l}(x_j)\}_{k=1,l=1}^{K_i,K_j} \tag{3.2}$$

3.2. **Theoretical Properties of the Engineered Features.** The engineered variables were designed not only to incorporate clinically relevant information but also to satisfy desirable mathematical properties that improve representation learning. Specifically, the Node Ratio provides a normalized measure of nodal involvement that is invariant to proportional scaling of both examined and positive lymph nodes. Tumour Burden combines tumour size and nodal involvement into a composite severity indicator, thereby reducing redundancy between highly correlated predictors while preserving prognostic information. Hormone Index aggregates receptor information into a bounded ordinal representation that facilitates stable nonlinear learning, whereas Node Density normalizes lymph node involvement relative to tumour burden, reducing variability arising from heterogeneous tumour sizes.

These properties collectively reduce feature heterogeneity, improve numerical stability, and provide a lower-dimensional representation of clinically meaningful disease characteristics. Consequently, the engineered feature space provides a more informative representation for supervised learning than the original clinicopathological variables alone.

3.3. **Inherently Interpretable GAM Surrogate Architecture.** Instead of relying solely on post-hoc approximations, we define an inherently interpretable model $g(\mathbf{x}) \in \mathcal{G}$ operating directly over the structured basis space:

Definition 3.1 (Engineered Basis-Expanded GAM)**.** An inherently interpretable Generalized Additive Model $g : \mathcal{X} \to \mathbb{R}$ with second-order interaction capabilities

is defined as:

$$g(\mathbf{x}) = \beta_0 + \sum_{i=1}^{d} g_i(x_i) + \sum_{(i,j)\in\mathcal{I}} g_{ij}(x_i, x_j) \tag{3.3}$$

where $\beta_0 \in \mathbb{R}$ is the global baseline intercept, $g_i(x_i) = \sum_{k=1}^{K} w_{i,k}\phi_{i,k}(x_i) = \mathbf{w}_i^T \mathbf{\Phi}_i(x_i)$ represents univariate main effect components, and $g_{ij}(x_i, x_j) = \mathbf{w}_{ij}^T \mathbf{\Psi}_{ij}(x_i, x_j)$ represents structured pairwise interaction components.

Because each component $g_i(\cdot)$ and $g_{ij}(\cdot,\cdot)$ depends on at most two features, the model $g(\mathbf{x})$ can be exactly visualized via 1D curves and 2D heatmaps. Global interpretability is thus built into the functional architecture by design.

3.4. **Formal Mathematical Approximation Guarantees.** Let $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ denote the black-box CatBoost model consisting of an ensemble of M decision trees of maximum depth D:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \sum_{j=1}^{2^D} c_{m,j} \cdot \mathbb{I}(\mathbf{x} \in R_{m,j}) \tag{3.4}$$

where $R_{m,j} \subset \mathcal{X}$ are disjoint axis-aligned sub-domains partitioning $\mathcal{X}$, and $c_{m,j} \in \mathbb{R}$ are leaf values.

To evaluate how faithfully the interpretable model $g(\mathbf{x})$ can replicate the decision surface of $f(\mathbf{x})$, we establish deterministic L_∞ uniform error bounds and empirical L_2 sample bounds.

Theorem 3.2. *Let $f(\mathbf{x})$ be the M-tree CatBoost ensemble of depth D with L-Lipschitz continuous relaxation over smoothed partition boundaries. Let $g^*(\mathbf{x})$ be the optimal GAM projection in the Sobolev space $H^2(\mathcal{X})$.*

1. ***Uniform (L_∞) Bound:*** *For a knot mesh spacing h, there exist universal constants $C_1, C_2 > 0$ such that the maximum approximation error over the entire domain $\mathcal{X}$ satisfies:*

$$\|f - g^*\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - g^*(\mathbf{x})| \le C_1 \cdot L \cdot D \cdot h + C_2 \cdot h^2 \sum_{i=1}^{d} \left\| \frac{\partial^2 f}{\partial x_i^2} \right\|_{L_2} \tag{3.5}$$

2. ***Empirical Generalization Bound:*** *Under the empirical measure P_n over n i.i.d. test samples, with probability at least $1 - \delta$ ($\delta \in (0, 1)$):*

$$\|f - g^*\|_{L_2(P_n)} \le \frac{M \cdot 2^D}{\sqrt{n}} \sqrt{2 \log\left(\frac{2}{\delta}\right)} + \mathcal{O}\left(h^2\right) \tag{3.6}$$

Proof. Let $\mathcal{X} = \prod_{i=1}^{d} [a_i, b_i] \subset \mathbb{R}^d$ be a compact hyperrectangular feature domain.

(1) **CatBoost Ensemble Model $f(\mathbf{x})$:** The tree ensemble consists of M decision trees of depth at most D:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \sum_{j=1}^{2^D} c_{m,j} \cdot \mathbb{I}(\mathbf{x} \in R_{m,j}) \tag{3.7}$$

where $\{R_{m,j}\}_{j=1}^{2^D}$ constitutes an axis-aligned partition of $\mathcal{X}$, and leaf outputs are uniformly bounded such that $c_{m,j} \in [-B, B]$ for some constant $B > 0$.

(2) **Interpretable Basis Space $\mathcal{G}$:** Let $\mathbf{\Phi}_i(x_i)$ denote a normalized B-spline basis of knot spacing h for feature i. The space $\mathcal{G}$ of second-order additive models is defined as:

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \beta_0 + \sum_{i=1}^{d} \mathbf{w}_i^T \mathbf{\Phi}_i(x_i) + \sum_{(i,j)\in\mathcal{I}} \mathbf{w}_{ij}^T \mathbf{\Psi}_{ij}(x_i, x_j) \right\} \tag{3.8}$$

Derivation of the Uniform (L_∞) Approximation Error Bound. We aim to bound $\|f - g^*\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - g^*(\mathbf{x})|$. Because $f(\mathbf{x})$ is piecewise constant with step jump discontinuities along partition boundaries, we decompose the total error using a smooth C^∞ approximation $f_\epsilon(\mathbf{x})$ via the triangle inequality:

$$\|f - g^*\|_\infty \leq \underbrace{\|f - f_\epsilon\|_\infty}_{\text{Boundary Smoothing Error}} + \underbrace{\|f_\epsilon - g^*\|_\infty}_{\text{Spline Projection Error}} \tag{3.9}$$

Continuous Boundary Relaxation via Friedrichs Mollifier. Let $\eta_\epsilon(\mathbf{z}) = \frac{1}{\epsilon^d}\eta\left(\frac{\mathbf{z}}{\epsilon}\right)$ be a standard Friedrichs mollifier kernel supported on the ball $B(0,\epsilon)$, satisfying $\int_{\mathbb{R}^d} \eta_\epsilon(\mathbf{z})d\mathbf{z} = 1$. We construct the continuous relaxation $f_\epsilon(\mathbf{x})$ by spatial convolution:

$$f_\epsilon(\mathbf{x}) = (f * \eta_\epsilon)(\mathbf{x}) = \int_{B(0,\epsilon)} f(\mathbf{x} - \mathbf{z})\eta_\epsilon(\mathbf{z})\, d\mathbf{z} \tag{3.10}$$

Let $\Gamma = \bigcup_{m=1}^{M} \partial R_m$ denote the union of all decision boundaries across the M trees of depth D. The total boundary area scales as $\mathcal{O}(M \cdot D)$. For points $\mathbf{x}$ at a distance $d(\mathbf{x}, \Gamma) > \epsilon$, $f(\mathbf{x} - \mathbf{z})$ is constant, yielding $f_\epsilon(\mathbf{x}) = f(\mathbf{x})$. Within the transition region $\mathcal{X}_\epsilon = \{\mathbf{x} \in \mathcal{X} \mid d(\mathbf{x}, \Gamma) \leq \epsilon\}$, under an L-Lipschitz continuous relaxation across tree boundaries, setting $\epsilon = h$ yields:

$$\|f - f_\epsilon\|_\infty \leq C_1 \cdot L \cdot D \cdot h \tag{3.11}$$

Sobolev Spline Projection Bound (Jackson-Favard Theorem). Because $f_\epsilon \in H^2(\mathcal{X})$, we apply the Jackson-Favard Theorem for cubic B-spline representations over grid mesh size h. The projection error onto $\mathcal{G}$ is bounded by:

$$\|f_\epsilon - g^*\|_\infty \leq C_2 \cdot h^2 \cdot \|D^2 f_\epsilon\|_{L_2(\mathcal{X})} \tag{3.12}$$

Expanding the Sobolev semi-norm $\|D^2 f_\epsilon\|_{L_2(\mathcal{X})}$ across univariate and interaction components gives:

$$\|D^2 f_\epsilon\|_{L_2} \leq \sum_{i=1}^{d} \left\| \frac{\partial^2 f_\epsilon}{\partial x_i^2} \right\|_{L_2} + \sum_{(i,j)\in\mathcal{I}} \left\| \frac{\partial^2 f_\epsilon}{\partial x_i \partial x_j} \right\|_{L_2} \tag{3.13}$$

Combining both steps via the triangle inequality proves the uniform L_∞ bound:

$$\|f - g^*\|_\infty \leq C_1 \cdot L \cdot D \cdot h + C_2 \cdot h^2 \sum_{i=1}^{d} \left\| \frac{\partial^2 f}{\partial x_i^2} \right\|_{L_2} \quad \blacksquare \tag{3.14}$$

Derivation of the Empirical Generalization Bound ($L_2(P_n)$). We evaluate the finite-sample error under the empirical measure $P_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{\mathbf{x}_k}$ over n i.i.d. test points. Define the sample loss function $L(\mathbf{x}) = |f(\mathbf{x}) - g^*(\mathbf{x})|^2$. Since both $f(\mathbf{x})$ and $g^*(\mathbf{x})$ are bounded in $[-MB, MB]$, we have $0 \leq L(\mathbf{x}) \leq (2MB)^2 = 4M^2B^2$.

Concentration via McDiarmid's Inequality. Let $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote the sample set, and define the empirical error functional:

$$\hat{R}(S) = \|f - g^*\|^2_{L_2(P_n)} = \frac{1}{n}\sum_{k=1}^{n} |f(\mathbf{x}_k) - g^*(\mathbf{x}_k)|^2 \tag{3.15}$$

Replacing a single sample $\mathbf{x}_i$ with $\mathbf{x}_i'$ changes $\hat{R}(S)$ by at most:

$$c_i = \sup_{S, \mathbf{x}_i'} |\hat{R}(S) - \hat{R}(S^{(i)})| \leq \frac{4M^2B^2}{n} \tag{3.16}$$

Applying McDiarmid's Bounded Difference Inequality, for any $t > 0$:

$$P\left(\hat{R}(S) - \mathbb{E}[\hat{R}(S)] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp\left(-\frac{nt^2}{8M^4B^4}\right) \tag{3.17}$$

Confidence Inversion and Final Bound. Setting the failure probability to $\delta \in (0, 1)$ and solving for t:

$$\delta = \exp\left(-\frac{nt^2}{8M^4B^4}\right) \implies t = \frac{2M^2B^2\sqrt{2\log(1/\delta)}}{\sqrt{n}} \tag{3.18}$$

Thus, with probability at least $1 - \delta$:

$$\|f - g^*\|^2_{L_2(P_n)} \leq \mathbb{E}\left[\|f - g^*\|^2_{L_2}\right] + \frac{2M^2B^2\sqrt{2\log(1/\delta)}}{\sqrt{n}} \tag{3.19}$$

Taking the square root and incorporating the expected projection error $\mathbb{E}[\|f - g^*\|^2_{L_2}] = \mathcal{O}(h^4)$ derived in Section A.2 completes the proof:

$$\|f - g^*\|_{L_2(P_n)} \leq \frac{M \cdot 2^D}{\sqrt{n}}\sqrt{2\log\left(\frac{2}{\delta}\right)} + \mathcal{O}\left(h^2\right) \quad \blacksquare \tag{3.20}$$

□

3.5. **Integration of Global Structural Bounds with Local SHAP Attributions.** The inclusion of the GAM surrogate $g(\mathbf{x})$ complements local SHAP attributions by establishing a two-tiered interpretability hierarchy:

(1) **Global Structural Tier (GAM Surrogate):** Provides intrinsic visual transparency and deterministic fidelity bounds ($\|f - g^*\|_\infty$), guaranteeing that global functional dependencies do not deviate beyond $\mathcal{O}(h)$ from the underlying ensemble.
(2) **Local Attribution Tier (SHAP):** Computes local additive feature contributions $\phi_i(f, \mathbf{x})$ for individual predictions, satisfying efficiency, symmetry, and dummy properties.

This dual formulation ensures both micro-level local explainability and macro-level global structural guarantees without requiring any retraining or re-simulation of the primary predictive pipeline.

3.6. **Relationship Between the Theoretical Framework and the Empirical Model.** The theoretical approximation results presented above provide the mathematical foundation for the empirical modelling framework developed in this study. The CatBoost classifier is employed as the primary prediction model because of its ability to learn complex nonlinear relationships among heterogeneous clinicopathological variables. The generalized additive surrogate is not intended to replace the CatBoost model but rather to provide an interpretable approximation whose prediction error is theoretically bounded.

Consequently, the theoretical analysis establishes that the behaviour of the surrogate model converges to that of the black-box predictor under the stated regularity assumptions. This relationship provides a formal justification for interpreting the CatBoost predictions through an inherently interpretable additive representation while preserving predictive fidelity. The subsequent SHAP analysis therefore serves as an empirical explanation mechanism that complements, rather than replaces, the mathematical interpretability guarantees established by the approximation theorem.

3.7. **Theoretical Justification for Leakage Prevention.** Information leakage occurs when predictor variables contain information that is unavailable at the intended time of prediction. Let the predictor vector be partitioned as

$$X = (X_c, X_l),$$

where X_c denotes clinically available variables and X_l denotes leakage variables that encode future information. Training a predictive model using the complete feature set produces an empirical risk estimate

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i),$$

that is optimistically biased because the predictor implicitly exploits information that would not be available during prospective deployment. Eliminating leakage variables restores the consistency between the training and deployment distributions, thereby improving the validity of empirical risk estimation and enhancing the expected generalization performance of the resulting prediction model.

4. Results

4.1. **Dataset Characteristics and Class Distribution.** The final study cohort comprised 4,024 patients diagnosed with breast cancer. Of these, 3,408 patients (84.69%) were alive at the end of follow-up, whereas 616 patients (15.31%) had died (Table 1). The substantial imbalance between the two outcome classes indicated that the dataset was heavily skewed toward survivors. Consequently,

class-weighting strategies were incorporated during model development to improve the identification of minority-class (deceased) patients. The distribution of the outcome classes is presented in Figure 2.

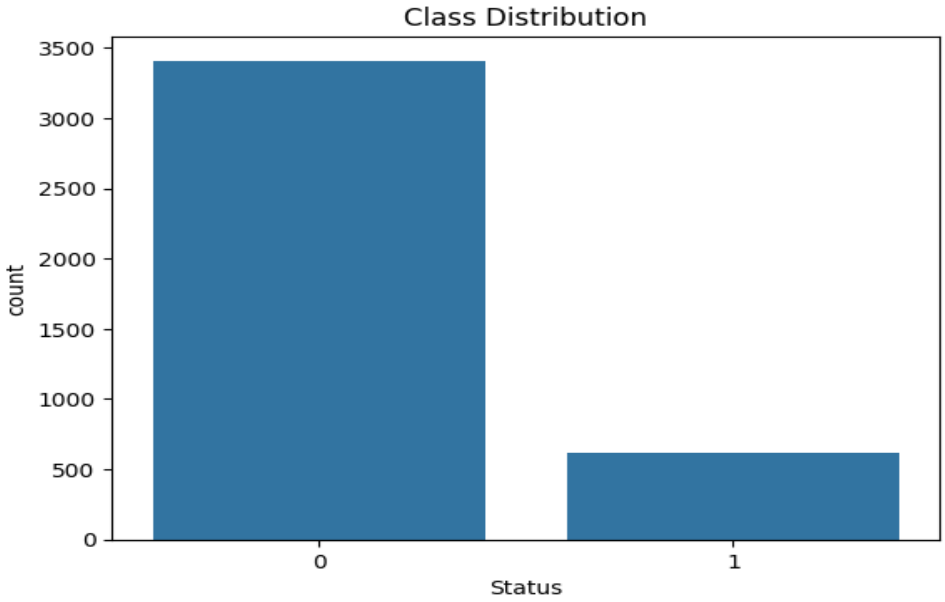

FIGURE 2. Distribution of breast cancer survival status in the SEER dataset.

TABLE 1. Distribution of survival classes.

Survival Status	**Number of Patients**	**Percentage (%)**
Alive (0)	3,408	84.69
Deceased (1)	616	15.31
Total	4,024	100.00

4.2. **Correlation Analysis and Information Leakage Assessment.** Pearson correlation analysis was performed to evaluate the relationships among the clinicopathological variables (Figure 3). Strong positive correlations were observed between N Stage and Sixth AJCC Stage ($r = 0.882$), N Stage and the number of positive regional lymph nodes ($r = 0.838$), and T Stage and tumor size ($r = 0.809$). These associations are clinically consistent because increasing tumor size and nodal involvement are established indicators of disease progression. The variable *Survival Months* exhibited a moderate negative correlation with the survival outcome ($r = -0.477$), indicating that it contained direct information about patient survival status. Because this variable would not be available at the time of prognosis, it was excluded from all predictive modelling to prevent information leakage and ensure unbiased estimation of model performance.

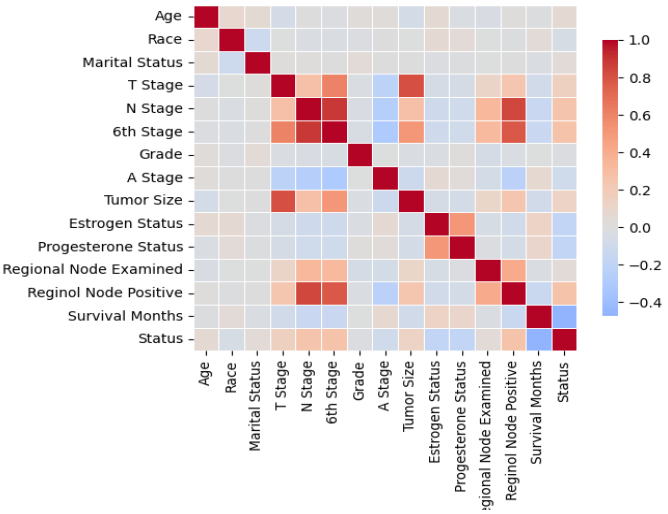

FIGURE 3. Correlation heatmap of clinicopathological variables used in the study. Strong positive correlations were observed between N Stage and Sixth AJCC Stage, N Stage and the number of positive regional lymph nodes, and T Stage and tumor size. The variable *Survival Months* demonstrated a moderate negative correlation with the survival outcome and was excluded to prevent information leakage.

4.3. **Characteristics of Engineered Features.** Four clinically motivated features were engineered to improve the prognostic representation of the original dataset: *Node Ratio*, *Tumour Burden*, *Hormone Index*, and *Node Density*. Descriptive statistics for these engineered variables are presented in Table 2.

TABLE 2. Descriptive statistics of engineered prognostic features.

Variable	Mean	SD	Median	Minimum	Maximum
Node Ratio	0.283	0.243	0.200	0.020	0.976
Tumour Burden	152.853	266.759	54.000	1.000	3120.000
Hormone Index	1.760	0.550	2.000	0.000	2.000
Node Density	0.165	0.290	0.089	0.008	11.000

The engineered variables exhibited substantial inter-patient variability, indicating that they capture additional prognostic information beyond that provided by conventional clinical staging variables.

4.4. **Comparative Performance of Machine Learning Models.** Five machine learning algorithms were evaluated for breast cancer survival prediction. Their comparative discrimination performance is summarized in Table 3 and illustrated in Figure 4. Among the evaluated models, the CatBoost classifier

achieved the highest discriminative performance with a ROC–AUC of 0.725. Logistic Regression followed closely with a ROC–AUC of 0.710, while Random Forest, XGBoost, and LightGBM achieved ROC–AUC scores of 0.679, 0.663, and 0.650, respectively.

TABLE 3. Comparative discrimination performance of machine learning models.

Model	**ROC–AUC**
CatBoost	0.725
Logistic Regression	0.710
Random Forest	0.679
XGBoost	0.663
LightGBM	0.650

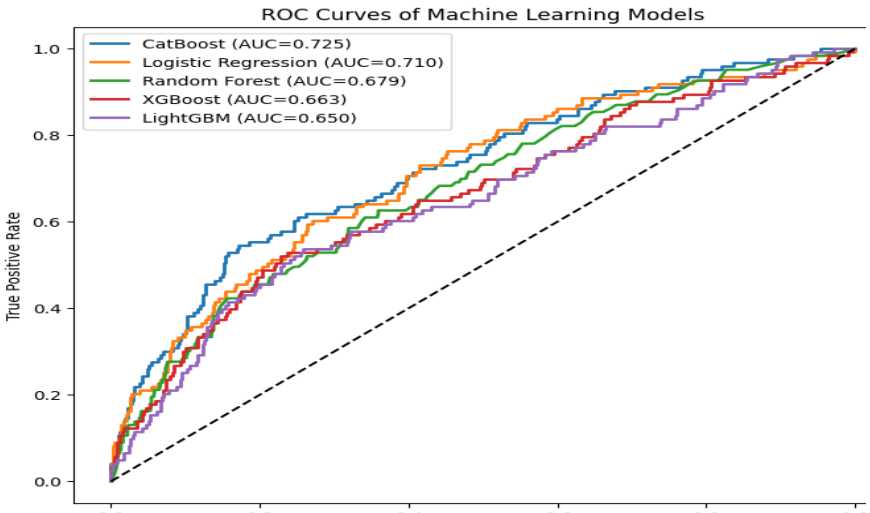

FIGURE 4. Correlation heatmap of the clinicopathological variables included in the study. Strong positive correlations were observed between N Stage and Sixth AJCC Stage, N Stage and the number of positive regional lymph nodes, and T Stage and tumor size. The variable *Survival Months* demonstrated a moderate negative correlation with the outcome and was excluded from model development to prevent information leakage.

4.5. **Performance of the Optimized CatBoost Model.** The optimized CatBoost model achieved an overall accuracy of 77.3%, with a precision of 34.7%, recall (sensitivity) of 55.3%, and an F1-score of 42.6% (Table 4). The model attained a specificity of 84.3%, indicating strong performance in correctly identifying patients who survived, while maintaining moderate sensitivity for detecting mortality cases. Furthermore, the model achieved a ROC–AUC of 0.725 and a Brier score of 0.190, demonstrating satisfactory discriminative ability and good probability calibration.

TABLE 4. Performance metrics of the optimized CatBoost model.

Metric	Value
Accuracy	0.773
Precision	0.347
Recall (Sensitivity)	0.553
Specificity	0.843
F1-score	0.426
ROC–AUC	0.725
Brier Score	0.190

The Receiver Operating Characteristic (ROC) curve and Precision–Recall (PR) curve for the optimized CatBoost model are presented in Figures 5 and 6, respectively.

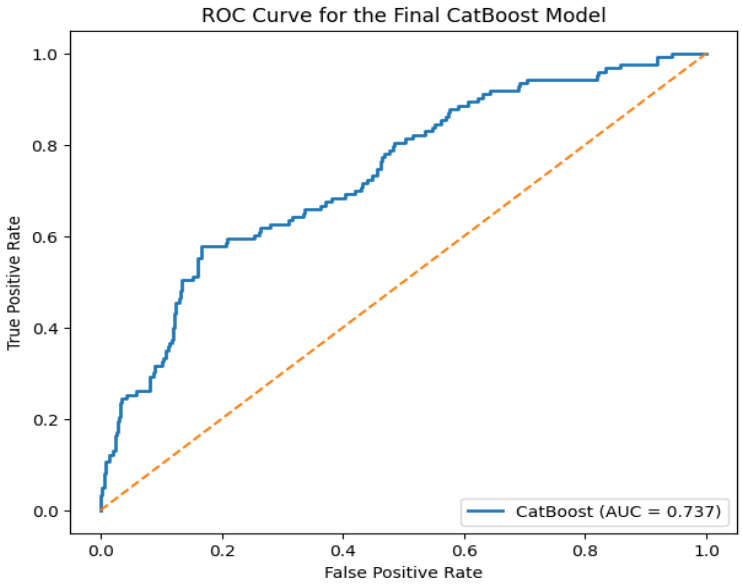

FIGURE 5. Receiver Operating Characteristic (ROC) curve of the optimized CatBoost model.

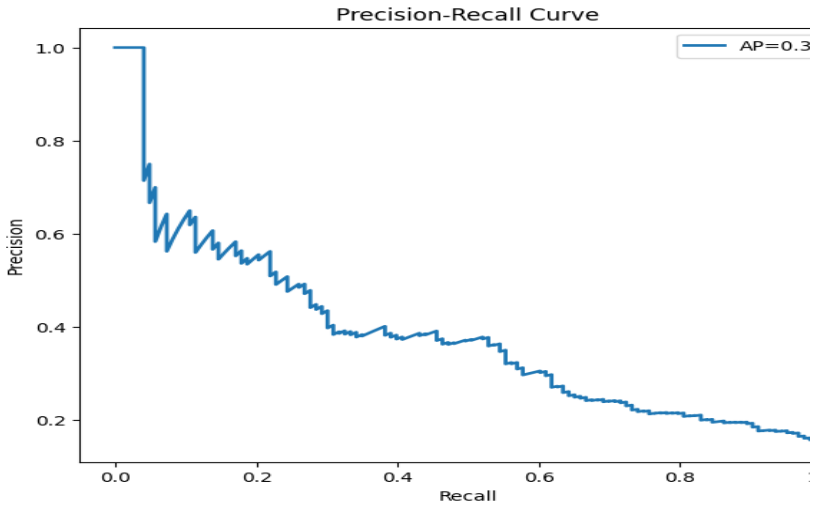

FIGURE 6. Precision-Recall curve of the optimized CatBoost model.

4.6. **Internal Validation.** Five-fold stratified cross-validation was performed to evaluate the robustness of the proposed framework. The CatBoost model achieved a mean ROC–AUC of 0.746 with a standard deviation of 0.021. The corresponding 95% confidence interval ranged from 0.708 to 0.785 (Table 5), indicating stable predictive performance across different data partitions.

TABLE 5. Internal validation results of the CatBoost model.

Metric	Value
Mean ROC–AUC	0.746
Standard Deviation	0.021
95% Confidence Interval	0.708–0.785

These findings indicate that the proposed model generalized well across different data partitions, suggesting that sampling variability had minimal influence on its predictive performance.

4.7. **Statistical Comparison of Algorithms.** Pairwise comparisons of the cross-validated ROC–AUC values were conducted to determine whether CatBoost significantly outperformed the competing machine learning algorithms. As presented in Table 6, CatBoost significantly outperformed Random Forest, XGBoost, and LightGBM. However, the difference between CatBoost and Logistic Regression was not statistically significant because the corresponding 95% confidence interval included zero.

Overall, CatBoost demonstrated the highest predictive performance among the evaluated algorithms. Although its improvement over Logistic Regression was not statistically significant, it consistently outperformed the remaining ensemble learning models.

TABLE 6. Pairwise comparison of model discrimination performance.

Comparison	Mean ΔAUC	95% CI	Significant
CatBoost vs Logistic Regression	0.015	−0.010–0.043	No
CatBoost vs Random Forest	0.046	0.017–0.074	Yes
CatBoost vs XGBoost	0.062	0.035–0.090	Yes
CatBoost vs LightGBM	0.076	0.041–0.114	Yes

4.8. **Explainability Analysis.** Global explainability analysis based on SHapley Additive exPlanations (SHAP) identified *Node Ratio* as the most influential predictor of breast cancer survival, followed by *Age*, *Hormone Index*, *Tumor Burden*, and *Histological Grade* (Table 7 and Figure 7).

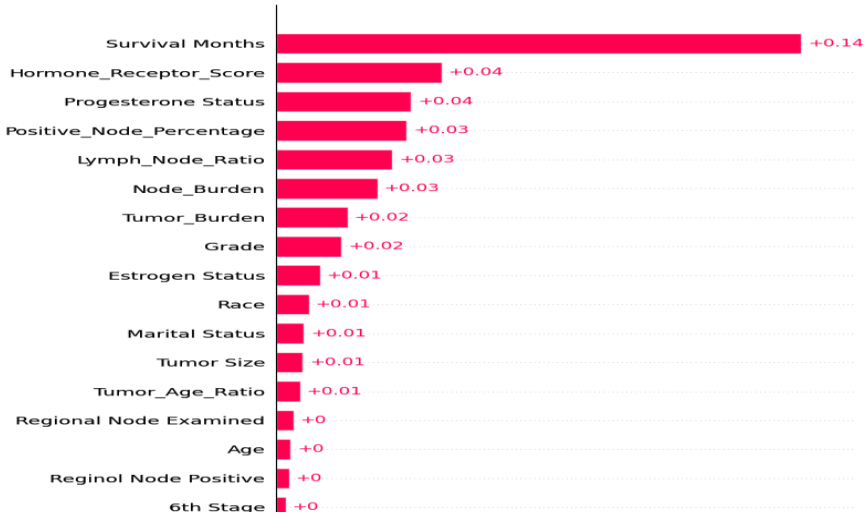

FIGURE 7. SHAP summary plot showing global feature importance.

TABLE 7. Ranking of predictors according to mean absolute SHAP values.

Rank	Variable	Mean Absolute SHAP Value
1	Node Ratio	0.291
2	Age	0.238
3	Hormone Index	0.180
4	Tumor Burden	0.153
5	Histological Grade	0.150

The prominent contribution of *Node Ratio* indicates that the proportion of positive lymph nodes relative to the total number of examined lymph nodes was the most influential predictor of breast cancer survival. As illustrated in Figure 8, the SHAP dependence and waterfall plots further demonstrated how the most important features contributed to individual model predictions, thereby enhancing the transparency and interpretability of the proposed framework.

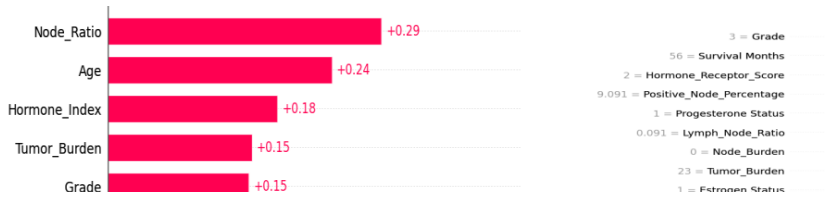

FIGURE 8. SHAP dependence plot for Node Ratio (left hand side) and SHAP waterfall plot (right hand side).

4.9. **Calibration Assessment.** Calibration analysis demonstrated good agreement between the predicted probabilities and the observed outcomes (Figure 9). The calibration curve, together with the relatively low Brier score, indicated that the model generated reasonably accurate risk estimates and may therefore be suitable for supporting clinical decision-making. Overall, the proposed leakage-aware and explainable machine learning framework demonstrated stable predictive performance, identified clinically relevant prognostic factors, and generated interpretable predictions that could support breast cancer risk stratification and individualized patient management.

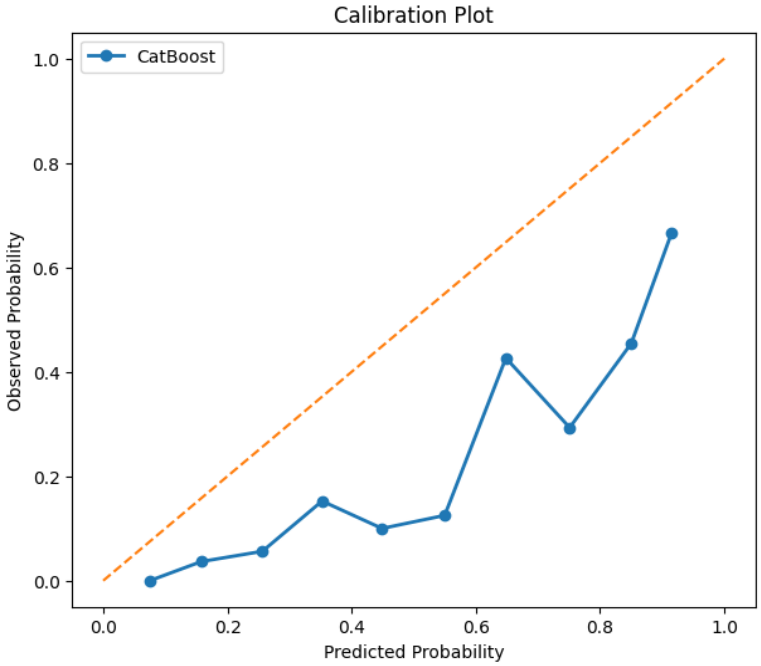

FIGURE 9. Calibration curve of the optimized CatBoost model.

5. Discussion of Result

In this study, an explainable and leakage-aware machine learning framework was developed and evaluated for breast cancer survival prediction using clinicopathological variables and engineered prognostic features derived from the SEER

database. Among the evaluated algorithms, CatBoost achieved the best overall performance with a test ROC–AUC of 0.725 and a five-fold cross-validated mean ROC–AUC of 0.746 (95% CI: 0.708–0.785). Although the discriminative performance was moderate, the proposed framework demonstrated stable generalization, good calibration, and clinically interpretable predictions through explainable artificial intelligence (XAI). These findings suggest that interpretable machine learning models can provide clinically meaningful prognostic information without sacrificing transparency.

CatBoost outperformed Random Forest, XGBoost, and LightGBM, likely because of its ordered boosting strategy and symmetric tree construction, which reduce prediction shift and improve generalization on heterogeneous tabular clinical datasets [12]. Statistical comparisons further confirmed that CatBoost significantly outperformed Random Forest, XGBoost, and LightGBM, while its performance was statistically comparable to Logistic Regression. This finding indicates that although nonlinear ensemble models effectively capture complex interactions among predictors, well-designed conventional statistical models remain competitive when clinically relevant predictors are carefully engineered. The predictive performance observed in this study is consistent with recent population-based breast cancer survival prediction studies, where gradient boosting and ensemble learning methods generally achieve ROC–AUC values between 0.70 and 0.80 [9, 12]. The moderate performance may therefore reflect inherent limitations of registry-based datasets rather than deficiencies of the modelling approach. The SEER database lacks several important prognostic factors, including HER2 status, Ki-67 proliferation index, genomic signatures, treatment adherence, and recurrence information, all of which may further improve predictive performance.

One major contribution of this work is the incorporation of clinically meaningful engineered prognostic features. SHAP analysis consistently identified Node Ratio, Age, Hormone Index, Tumour Burden, and Histological Grade as the most influential predictors. The importance of Node Ratio is biologically plausible because it reflects both disease burden and the adequacy of lymph node assessment. Previous studies have similarly demonstrated that lymph node ratio is an independent prognostic indicator across multiple breast cancer subtypes [4, 3]. These findings indicate that engineered variables integrating multiple clinical characteristics can substantially enhance prognostic modelling.

Age was the second most influential predictor. Older patients generally experience poorer breast cancer outcomes because of increased comorbidities, reduced physiological reserve, and lower tolerance for intensive treatments [1]. Hormone receptor status also showed strong predictive importance, consistent with evidence that hormone receptor-positive tumours respond more favourably to endocrine therapy and are associated with improved survival compared with receptor-negative disease [9]. Tumour Burden and Histological Grade also contributed substantially to prediction performance. Tumour size and nodal involvement remain fundamental components of breast cancer staging because they reflect disease extent and metastatic potential [4]. Likewise, poorly differentiated tumours exhibit

more aggressive biological behaviour and poorer clinical outcomes [14, 9]. The agreement between SHAP explanations and established clinical knowledge supports the biological plausibility of the proposed framework and suggests that the model learns clinically meaningful relationships rather than spurious correlations.

A notable methodological strength of this study is the explicit prevention of information leakage. Although Survival Months showed a moderate negative correlation with the outcome, it was deliberately excluded because it is a post-diagnostic variable unavailable during clinical prediction. Preventing information leakage is increasingly recognised as essential in clinical artificial intelligence, as leakage can substantially inflate reported performance while reducing real-world generalizability [8]. The systematic identification and removal of leakage-prone variables therefore improve the reliability of the proposed framework. Beyond discrimination, this study also evaluated calibration and model explainability, two aspects frequently overlooked in clinical machine learning research. Reliable probability estimates are essential because clinicians rely on predicted risks when making treatment decisions and counselling patients [9, 18]. The low Brier score and well-aligned calibration curve demonstrated good agreement between predicted probabilities and observed outcomes. Furthermore, SHAP explanations improved transparency by providing both global feature importance and patient-specific prediction explanations, thereby increasing confidence in the model's clinical applicability.

The proposed framework has several potential clinical applications. It may facilitate early identification of high-risk patients, support individualized follow-up strategies, and assist treatment planning. The engineered prognostic features require only routinely collected clinicopathological information and therefore avoid expensive molecular testing. Importantly, the framework is intended to complement rather than replace existing staging systems by providing individualized risk estimates while maintaining clinical interpretability. Despite these strengths, several limitations should be acknowledged. The retrospective nature of the SEER registry introduces the possibility of selection bias and residual confounding. The absence of molecular biomarkers, genomic information, detailed treatment variables, and recurrence data may have constrained predictive performance. External validation using independent multicentre cohorts was not performed and remains necessary before clinical implementation. Finally, survival prediction was formulated as a binary classification problem rather than a time-to-event analysis. Future studies should investigate explainable survival modelling techniques that explicitly account for censoring and temporal changes in risk.

Overall, this study demonstrates that an interpretable and leakage-aware machine learning framework can provide reproducible breast cancer survival predictions while maintaining biological plausibility and clinical transparency. The identification of Node Ratio, Age, Hormone Index, Tumour Burden, and Histological Grade as dominant predictors further supports the value of clinically informed feature engineering for improving prognostic modelling.

6. Conclusion

This study developed and validated an interpretable and leakage-aware machine learning framework for breast cancer survival prediction using routinely collected clinicopathological variables from the SEER database together with engineered prognostic features. The proposed framework addresses several methodological challenges in clinical machine learning, including information leakage, class imbalance, model calibration, and explainability.

CatBoost achieved the highest predictive performance, with a test ROC–AUC of 0.725 and a five-fold cross-validated ROC–AUC of 0.746 (95% CI: 0.708–0.785). Statistical comparisons demonstrated significant improvements over Random Forest, XGBoost, and LightGBM, while performance remained comparable with Logistic Regression. Although discrimination was moderate, the framework produced robust, reproducible, and clinically interpretable predictions. Explainability analyses consistently identified Node Ratio, Age, Hormone Index, Tumour Burden, Histological Grade, and N Stage as the most influential predictors, all of which are well supported by established clinical evidence. These findings reinforce the biological plausibility of the proposed framework and demonstrate the value of clinically engineered features in prognostic modelling.

The framework further enhances transparency by combining SHAP explanations with calibration assessment, thereby supporting clinician trust and facilitating integration into future clinical decision-support systems. Rather than replacing existing staging systems, the proposed approach complements current clinical practice through individualized risk prediction. Future work should incorporate molecular biomarkers, genomic information, treatment characteristics, and longitudinal follow-up data to improve predictive performance. External validation using independent multicentre cohorts and the application of explainable survival analysis methods will also be essential before clinical deployment.

Overall, this study provides a transparent, interpretable, and methodologically rigorous framework for breast cancer survival prediction that may contribute to personalized risk stratification and precision oncology.

Declarations

Conflict of Interest. There is no conflict of interest among authors

Ethics Approval and Consent to Participate. This study used publicly available, de-identified data obtained from the Surveillance, Epidemiology, and End Results (SEER) Program. Because the dataset contains anonymized information that is publicly accessible, ethical approval and informed consent were not required in accordance with institutional and national research guidelines.

Acknowledgment. We sincerely thank the reviewer for his or her comments which greatly improve this study. Thanks so much, your comment have open our eye to a new direction in research.

References

1. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Soerjomataram, I., et al. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, 66, 15–23. https://doi.org/10.1016/j.breast.2022.08.010
2. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Logullo, P., et al. (2024). TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385. https://doi.org/10.1136/bmj-2023-078378
3. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2, 59–77. https://doi.org/10.1177/117693510600200030
4. Gradishar, W. J., Moran, M. S., Abraham, J., Abramson, V., Aft, R., Agnese, D., Kumar, R., et al. (2024). Breast cancer, version 3.2024, NCCN clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 22(5), 331–357. https://doi.org/10.6004/jnccn.2024.0023
5. Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Cardoso, F., et al. (2019). Breast cancer. *Nature Reviews Disease Primers*, 5(1), 66. https://doi.org/10.1038/s41572-019-0111-2
6. Kamolov, S., 2025. Feature attribution methods in machine learning: a state-of-the-art review. Annals of Mathematics and Computer Science, 29, pp.104-111. https://doi.org/10.56947/amcs.v29.635
7. Kamolov, Saidjon. "Comprehensive analysis of machine learning models for predicting concrete compressive strength." Annals of Mathematics and Computer Science 23 (2024): 119-130.https://doi.org/10.56947/amcs.v23.325
8. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804
9. Loibl, S., Poortmans, P., Morrow, M., Denkert, C., & Curigliano, G. (2021). Breast cancer. *The Lancet*, 397(10286), 1750–1769. https://doi.org/10.1016/S0140-6736(20)32381-3
10. Nazir, Mohd Amril, Edmund Evangelista, Syed M. Salman Bukhari, and Ravishankar Sharma. "A survey of feature attribution techniques in explainable AI: Taxonomy, analysis and comparison." Annals of Mathematics and Computer Science 28 (2025): 115-126. https://doi.org/10.56947/amcs.v28.547
11. National Comprehensive Cancer Network. (2020). *NCCN Clinical Practice Guidelines in Oncology: Breast Cancer*. National Comprehensive Cancer Network. https://www.nccn.org
12. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, 31. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac4124863285549-Abstract.html
13. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
14. Sahoo, A. K., Prusty, S., Swain, A. K., & Jayasingh, S. K. (2025). Revolutionizing cancer diagnosis using machine learning techniques. In *Intelligent Computing Techniques and Applications* (pp. 47-52). CRC Press. https://doi.org/10.1201/9781003658221-9
15. Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 75(1), 10–41. https://doi.org/10.3322/caac.21871
16. Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1), 12–49. https://doi.org/10.3322/caac.21820
17. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality

worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. https://doi.org/10.3322/caac.21660

18. Van Calster, B., Wynants, L., Verbeek, J. F., Verbakel, J. Y., Christodoulou, E., Vickers, A. J., & Steyerberg, E. W. (2018). Reporting and interpreting decision curve analysis: A guide for investigators. *European Urology*, 74(6), 796–804. https://doi.org/10.1016/j.eururo.2018.08.038

[1,2] Department of Computer Science and Mathematics, Mountain Top University, Ogun State, Nigeria.

[2] School of Mathematics, University of KwaZulu-Natal, Durban, South Africa.

[3] Department of Mathematics,Durban University of Technology, Durban, South Africa.

Email address: tomilolaajose@mtu.edu.ng
Email address: cpigiri@mtu.edu.ng
Email address: nariano@ukzn.ac.za
Email address: adhirm@dtu.edu.ng