Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 5 (2026) 1371 - 1380 e-ISSN: 2580-0760

Quantifying Target Leakage in SEER Breast Cancer Mortality Prediction

Sestri Novia Rizki[1], Vani Maharani Nasution[2], Eko Amri Jaya[3], Verra Budhi Lestari[4], Elfina Maulid[5]
[1,3,4,5]Universitas Media Nusantara Citra, Jakarta, Indonesia
[2]Universitas Jakarta Internasional, Jakarta, Indonesia

*noviasestri@gmail.com

Abstract

Recent machine learning studies on the SEER Breast Cancer dataset report 95–98% accuracy (2022–2025), but their methods are rarely examined. Because the dataset is imbalanced (84.7% Alive), accuracy alone is a poor discrimination metric — a trivial always-Alive classifier scores near 85% — so this audit quantifies leakage inflation on ROC-AUC, computed identically across 12 models. We measure how much performance stems from target leakage via Survival Months, which encodes retrospective survival time unavailable at prediction. Undetected, such leakage could lead clinicians to overtrust a model's risk estimates, risking inappropriate care decisions. We train 12 models under three conditions: a) naïve, with Survival Months and a non-stratified split; b) leakage-free with stratified evaluation; and c) fixed-horizon (36- and 60-month) binary mortality, excluding censored cases. Under Condition A, the best ROC-AUC is 0.889 (accuracy 0.899). Removing leakage yields a consistent AUC ceiling of 0.74–0.77 across all 12 models — an inflation of +0.07 to +0.20 AUC points. SHAP shows Survival Months dominates feature importance by 3.0–4.3× under leakage; imbalance-handling methods (BorderlineSMOTE, class weighting) raise minority-class F1 but lower mean Brier Skill Score to −0.585 and −0.597. The contribution is the audit methodology itself — a reusable protocol for detecting leakage-driven inflation in registry-based clinical prediction.

Keywords: breast cancer mortality prediction; class imbalance; machine learning audit; SEER dataset; SHAP analysis

How to Cite: S. N. . Rizki, V. M. Nasution, . E. A. . Jaya, V. B. . Lestari, and Elfina Maulid, "Quantifying Target Leakage in SEER Breast Cancer Mortality Prediction", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 5, pp. 1371–1380, Oct. 2026, doi: 10.29207/resti.v10i5.8105. *Permalink/*DOI: https://doi.org/10.29207/resti.v10i5.8105

Received: June 18, 2026
Accepted: September 5, 2026
Available Online: October 7, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Breast cancer is the most diagnosed malignancy in women worldwide and a leading cause of cancer death in all sexes [1], [2]. GLOBOCAN 2022 estimates about 2.3 million new diagnoses and 670,000 deaths in a single year [3]. Accurate prognosis is essential for treatment planning, resource allocation, and patient counseling [4], [5]. With the growing availability of large public cancer registries such as SEER [6], many studies now apply supervised machine learning (ML) to predict breast cancer survival [7], [8], often reporting striking performance.

From 2022 to 2025, several machine learning studies using the SEER breast cancer dataset have reported high predictive performance for survival-related outcomes [1], [9], [10]. Manikandan et al. used ensemble and tree-based classifiers and reported 98% accuracy for the Decision Tree model using both train–test split and cross-validation [9]. Several machine learning studies on breast cancer classification have reported strong performance for logistic regression [11], SVM [12], [13], and neural network-based models [14], often achieving accuracies above 87% [15-17].

A closer look shows these figures are likely products of compounded methodological flaws, not genuine predictive power. Three related problems have been largely overlooked. The most serious is *target leakage* via the `Survival Months` feature — broadly, a model gains access, at training time, to information that would not be available at the moment a real prediction is actually made, typically because the leaking feature is generated by the same process as, or is causally downstream of, the outcome being predicted. Many studies keep survival duration as an input when training a classifier to predict the binary outcome `Status` (Alive/Dead).

We concentrate the audit on this single feature rather than distributing attention across all clinical covariates because it is structurally distinct from the rest: stage, grade, and hormone-receptor status are established at diagnosis and are legitimate, if imperfect, prognostic factors, whereas Survival Months is only fully

determined once the patient's outcome is known — it is generated by the same death-or-censoring process as the label itself, making it a textbook case of leakage rather than a merely correlated risk factor.

This is direct target leakage; training uses diagnosis information unavailable at prediction time because survival duration is known only afterward. A model that infers death from survival time is not performing true prognosis. A landmark survey of ML-based science found leakage in 294 papers across 17 fields [18], sometimes inflating reported performance by over 12 percentage points; similar leakage-driven inflation has been systematically documented in neuroimaging [19], [20].

The second problem is using overall accuracy as the main metric despite severe class imbalance. In this dataset, 84.7% of patients are Alive and 15.3% are Dead. With such skew, a trivial classifier that always predicts "Alive" would achieve about 85% accuracy without learning anything clinically meaningful, a known issue called the accuracy paradox [21].

The third problem is ignoring the time-to-event nature of registry data. The binary label `Status` (Alive/Dead) is set at registry extraction rather than at a defined follow-up time. Patients labeled "Alive" can have very different follow-up durations, and some may die soon after the snapshot. This conflates cross-sectional status with long-term prognosis, ignores censoring and leaves the modeling target clinically undefined [22], [23].

Prior work has addressed these issues separately. One study reformulated the SEER dataset as fixed-horizon binary classification at 3- and 5-year outcomes [24], comparing L1-regularized logistic regression with a feedforward neural network while correctly excluding censored observations. The logistic model achieved ROC-AUCs of 0.78 (3 years) and 0.75 (5 years), far below the 95–98% accuracy reported in leakage-prone studies.

Our Condition C AUCs (0.767 at 36 months, 0.740 at 60 months; Section 3.6) closely replicate this leakage-free benchmark despite differences in model family, feature engineering, and horizon definition, corroborating that an AUC of roughly 0.74–0.78 is a genuine, reproducible ceiling for this dataset with standard clinical features rather than an artifact of any single pipeline. Explainability-focused research has also applied SHAP (SHapley Additive exPlanations) to identify clinically meaningful predictors in breast cancer models [25], [26].

This paper addresses that gap with a controlled, reproducible empirical audit. We deliberately reuse the same SEER Breast Cancer dataset as the audited 2022–2025 studies rather than a different cohort: holding the data fixed isolates methodology as the source of the reported performance gap, rather than confounding it with differences in patient population, feature availability, or class balance across datasets; it also lets our leakage-free benchmark serve as a direct point of comparison for the very papers it audits.

We train and evaluate 12 classifiers under three conditions: naïve replication using `Survival Months` with a non-stratified split, leakage-free cross-sectional binary classification with stratified evaluation and fixed-horizon prediction at 36 and 60 months, excluding censored observations. Each condition is evaluated with a comprehensive metric suite plus SHAP feature importance and paired statistical tests [27]. A secondary contribution quantifies a previously unreported discrimination–calibration trade-off on this dataset.

The main contributions of this work are a controlled three-condition empirical audit that isolates leakage from algorithmic and evaluation confounders; SHAP-based evidence that feature importance changes fundamentally when leakage is removed; and a leakage-aware benchmark that offers a realistic performance ceiling for this dataset for positioning future work.

2. Methods

This controlled empirical audit applies the same dataset, 12-model zoo, and evaluation protocol across three experimental conditions that differ only in prediction targets and feature sets. The research pipeline follows six sequential stages in two phases, as shown in Figure 1.

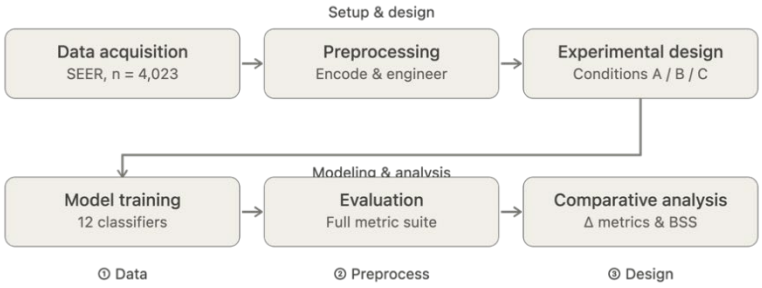


Figure 1. Research Pipeline

2.1 Dataset and Preprocessing Pipeline

The dataset is the publicly available SEER Breast Cancer Dataset (SEER November 2017 update), accessed via IEEE DataPort [6]; no new data were collected from human subjects, and all records were de-identified prior to public release under the data-use

terms of the original registry. As such, this analysis does not require additional institutional ethical review beyond the approvals already governing the public release of the dataset.

Four data quality corrections are applied before modeling: an empty column is dropped; trailing whitespace is removed from the `Race` and `T Stage` column names; `Reginol Node Positive` is renamed to `Regional Node Positive` to fix a typo; and one fully duplicated row is removed. The resulting dataset contains 4,023 patients and 15 variables. The binary outcome `dead` (0 = Alive, 1 = Dead) is derived from `Status`, with 84.69% Alive and 15.31% Dead, indicating class imbalance that affects metric choice and interpretation of accuracy.

Preprocessing steps fall into two categories that receive different treatment with respect to the train-test split. Deterministic recodings do not depend on any sample-wise statistic and are therefore identical regardless of split; for convenience these are applied once to the full 4,023-patient dataset. Ordinal clinical variables (T Stage, N Stage, 6th Stage, Grade) are integer-encoded to preserve clinical order (e.g., Grade I–IV → 1–4), avoiding the artificial equidistance of one-hot encoding for ranked variables.

Binary variables (A Stage, Estrogen Status, Progesterone Status) are encoded as 0/1. Categorical variables Race (three levels) and Marital Status (five levels) are one-hot encoded, with one reference category dropped for each to avoid perfect multicollinearity.

The four continuous variables — Age, Tumor Size, Regional Node Examined, and Regional Node Positive, are kept in raw form for tree-based models, which are invariant to monotonic feature transformations. For linear and distance-based models (Logistic Regression L1/L2, k-Nearest Neighbors, Naïve Bayes, SVM, MLP), we build a second feature matrix where all continuous variables are standardized with StandardScaler to reduce the impact of extreme values on gradient-based optimization.

Unlike the deterministic recodings above, StandardScaler is a fitted transformation: its mean and scale parameters are estimated exclusively on the training partition (or the training fold within five-fold cross-validation) and then applied to transform the corresponding held-out test set (or validation fold), never the reverse. The same training-only-fitting rule governs the resampling methods (BorderlineSMOTE, class weighting) in Section 3.5 and model fitting itself, all of which are confined to the training partition or CV training fold. This yields two feature matrices: X_raw for tree-based/boosting models and X_scaled for linear and distance-based models.

We compute one engineered feature, `node_positive_ratio`, as `Regional Node Positive` divided by `Regional Node Examined`, clipped to [0, 1]. Because both counts are known at pathological staging, this is a valid diagnosis-time feature and is included in all conditions and model families for fair comparison.

2.2 Experimental Design

The audit uses three experimental conditions, each targeting a specific source of performance inflation. Condition A (Naïve replication) includes `Survival Months` and all clinical variables as features, uses the cross-sectional binary outcome `dead`, and applies a non-stratified train-test split (`test_size = 0.2`, `random_state = 42`). This condition intentionally reproduces the most common experimental protocol used in 2022–2025 studies on this dataset.

Condition B is leakage-removed and cross-sectional. The `Survival Months` feature is entirely excluded; the outcome remains the cross-sectional `dead` label. Both the holdout split and the five-fold cross-validation on the training set are stratified by class label, preserving the ~85/15 class ratio in each fold.

Condition C is a fixed-horizon setup that fixes both leakage and the temporal validity of the prediction target, evaluated at two clinically motivated horizons: H = 36 and H = 60 months. For each horizon, the binary outcome event_H is derived from (Status, Survival Months) as follows: event_H = 1 if Status = Dead and survival ≤ H; event_H = 0 if Status = Dead and survival > H (death after the horizon) or if Status = Alive and survival ≥ H (confirmed survival beyond H).

In plain terms: a patient alive with documented follow-up at or beyond H is a confirmed non-event, and a patient who died after H is also coded as a non-event at that horizon — their death simply falls outside the H-month prediction window; only deaths occurring at or before H are labeled positive.

Patients who are alive with follow-up < H are excluded (censored), since their outcome at H is unknown. This yields effective cohorts of 3,956 patients for H = 36 months (67 excluded; event rate 5.94%) and 3,270 patients for H = 60 months (753 excluded; event rate 14.01%). The same stratified split and five-fold cross-validation as in Condition B is applied separately to each horizon cohort. Reporting two horizons additionally serves as a basic sensitivity check on horizon choice: both yield AUC values in the same 0.74–0.77 range as Condition B (Section 3.6), indicating the leakage-free performance ceiling is not an artifact of one arbitrary cutoff.

We further tested the censoring rule itself by re-running the H = 36 months cohort under an alternative, non-standard assignment that retains censored-alive patients and labels them as non-events instead of excluding them. This alternative reduces ROC-AUC for both Logistic Regression (0.766 → 0.749) and Gradient Boosting (0.750 → 0.735) relative to the exclusion-based cohort reported here, confirming that excluding patients with unknown horizon-outcome status is not a

convention of choice but a necessary step to avoid diluting the signal with mislabeled censored cases.

2.3 Model Zoo

For each condition–variant combination, twelve supervised classifiers are trained: Logistic Regression with L1 (`solver='liblinear'`) and L2 regularization; k-Nearest Neighbors (k=5); Gaussian Naïve Bayes; a single Decision Tree; an RBF SVM (`probability=True`); a 200-tree Random Forest; AdaBoost; Gradient Boosting; XGBoost (`eval_metric='logloss'`); LightGBM; and a single-hidden-layer MLP with 32 units. All models use `random_state=42`.

All models use random_state=42. The zoo excludes CatBoost and the Explainable Boosting Machine (EBM). CatBoost's principal advantage — native ordered-statistics handling of high-cardinality categorical features — offers little benefit here, since the two categorical variables (Race, Marital Status) have only three and five levels and are already one-hot encoded without loss; including it would also require a CatBoost-specific preprocessing branch that breaks the identical-feature-access principle below.

EBM is a glass-box generalized additive model with its own built-in, exact explanations, which sit awkwardly next to this paper's post-hoc SHAP analysis of black-box tree ensembles (Section 3.3); comparing a self-explaining model against SHAP-explained ones would need a separate interpretability framework outside this paper's scope. Both are reasonable extensions for future work.

Fair treatment across the twelve models is enforced at three levels: identical data — the same train/test split, and the same five-fold cross-validation folds, is reused for every model within a condition, so no model sees an easier or harder partition; identical feature access — every model receives the same feature set for its family (X_raw for tree-based/boosting models, X_scaled for linear and distance-based models, per Section 2.1), never a customized subset; and identical configuration effort — a single random_state=42 and library-default hyperparameters are applied uniformly, so no model benefits from a tuning budget the others did not receive.

No extensive hyperparameter tuning is performed by design. Allowing different tuning efforts across conditions would confound optimization depth with methodological correctness. We use library default hyperparameters throughout and acknowledge this as a scope limitation.

2.4 Evaluation Metrics and Statistical Tests

SHAP values are computed with shap.TreeExplainer on the output scale, using feature_perturbation= "tree_path_dependent" as the background distribution. All experiments use Python, scikit-learn, shap, xgboost, lightgbm, and imbalanced-learn.

Bootstrap confidence intervals 1,000 resamples, are computed for Accuracy, f1_dead, ROC-AUC, and PR-AUC. DeLong's test is implemented implemented in Python following the fast DeLong algorithm and McNemar's test.

2.5 Evaluation Metrics and Statistical Tests

Each model is evaluated on the held-out test set using 12 metrics. Discrimination metrics are Accuracy, Precision, Recall/Sensitivity, Specificity, F1-score for the Dead class, Balanced Accuracy, G-mean, ROC-AUC, PR-AUC, Matthews Correlation Coefficient, and Cohen's κ. Calibration is assessed with Brier Score and Expected Calibration Error.

Because target prevalence differs across conditions, raw Brier scores are not directly comparable. We therefore use the Brier Skill Score, where Brier_baseline = $\bar{p}(1-\bar{p})$ and $\bar{p}$ is the test-set positive-class prevalence. BSS > 0 indicates better calibration than a constant-prevalence predictor; BSS ≤ 0 indicates no calibration gain over predicting the base rate for every patient.

Cross-condition performance differences are evaluated with three statistical tests. DeLong's test compares paired AUCs on the same test fold, and McNemar's test assesses paired error-rate differences. Bootstrap confidence intervals (1,000 resamples) are computed for Accuracy, `f1_dead`, ROC-AUC, and PR-AUC. SHAP values are calculated for the four tree-based ensemble models under Conditions A and B to characterize how feature importance shifts after leakage removal.

3. Results and Discussions

Before presenting model results, it is important to understand how class distribution changes across the three experimental target formulations, as this directly affects performance metric interpretation. Figure 2 shows the proportions of positive-class patients (Dead / event) and negative-class patients (Alive / survived horizon) for each target definition.

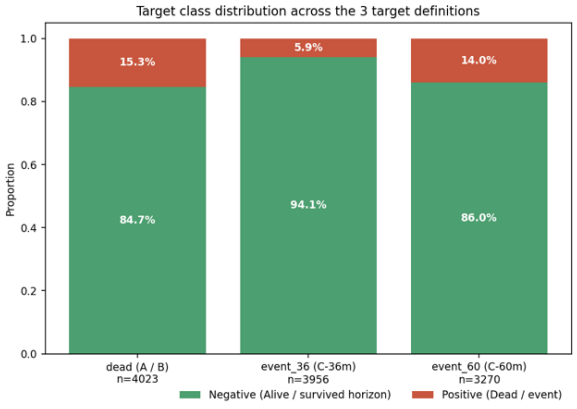


Figure 2. Target class distribution across the three target formulations

For Conditions A and B (the cross-sectional `dead` label), the positive rate is just 15.3%, creating strong

class imbalance that makes accuracy misleading. A trivial model that always predicts "Alive" would still achieve about 85% accuracy while having no discriminative power.

For Condition C at 36 months, the event rate drops to 5.9% because 67 censored observations are excluded and the horizon is shifted to a time when fewer patients have died. This severe imbalance makes C-36m the hardest of the four conditions and invalidates raw Brier Score comparisons across conditions, since a lower event rate trivially lowers the Brier baseline.

3.1 Performance Under the Leakage-Inclusive Baseline (Condition A)

Under Condition A, the 12-model zoo attains accuracies in line with 2022–2025 results on this dataset. Gradient Boosting performs best (accuracy 0.8994, ROC-AUC 0.8888), while the Decision Tree reaches 0.8311 accuracy and 0.7168 ROC-AUC. The non-stratified train-test split produces a test dead-class prevalence of 16.89%, slightly above the training set's 14.92%.

Within Condition A, the accuracy paradox appears clearly when all metrics are considered. Gradient Boosting reaches 0.8994 accuracy but only 0.6553 F1 for the Dead class, missing about one third of positive cases. KNN attains 0.865 accuracy yet an F1(dead) of 0.452, identifying fewer than half of deceased patients. This gap between accuracy and minority-class F1 is not incidental; it is the structural consequence of an 84.7/15.3 class split, where predicting the majority class correctly on most samples is sufficient to sustain high accuracy while minority-class recall remains poor.

Under Condition A, the Decision Tree's calibration is revealing. Its Brier Skill Score is −0.203, indicating that even with the leakage variable `Survival Months`, its probability estimates are worse than always predicting the base-rate prevalence. This illustrates single trees' tendency to produce extreme, poorly calibrated probabilities from pure leaves and shows that the Decision Tree is both a weak discriminator and a poor calibrator.

3.2 Quantifying Leakage Inflation: Condition A vs. Conditions B and C

Figure 3 presents the main empirical result: the AUC inflation from including `Survival Months` in the feature set, shown for all three leakage-free targets and all 12 models. Each model has three bars—Δ(A−B) in red, Δ(A−C36m) in orange, and Δ(A−C60m) in purple. All 12 models show positive inflation on all three comparisons; the y-axis starts at 0, so there are no negative deltas.

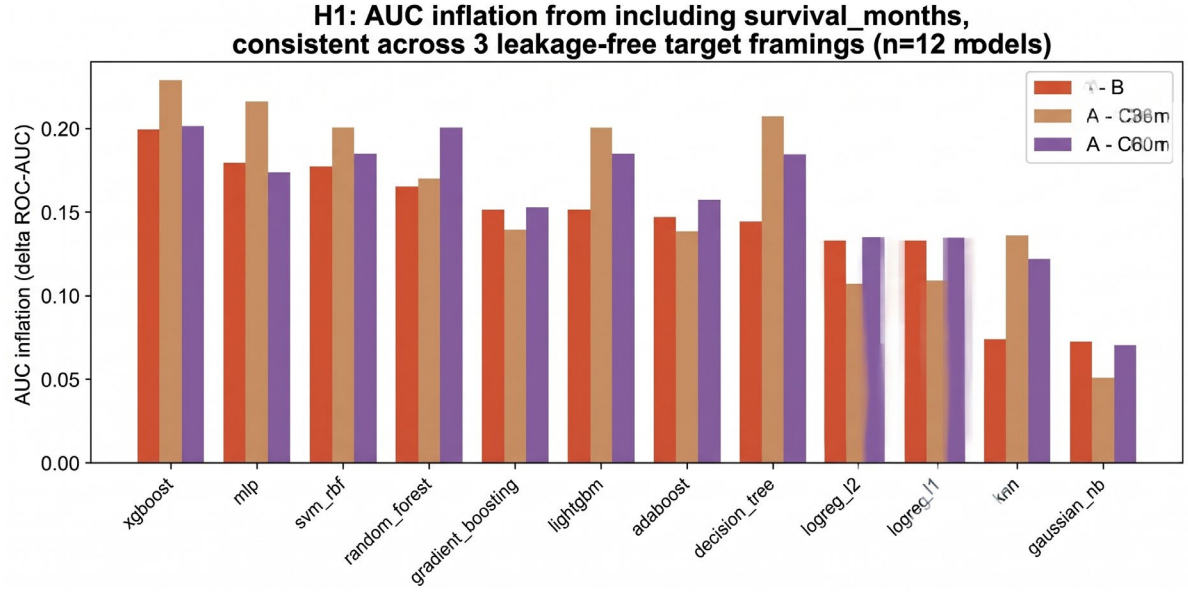


Figure 3. AUC inflation from including `Survival Months`

Under Condition B, the best ROC-AUC (0.746) is achieved by Logistic Regression. The top F1(dead) result is Logistic Regression (L2) with BorderlineSMOTE: F1(dead) = 0.4142, accuracy = 0.7329, AUC = 0.7415. The lower accuracy than in Condition A does not indicate worse modeling but more honest evaluation, as the model is no longer rewarded for exploiting retrospective time-encoding and is properly penalized for its true minority-class performance.

3.3 Feature Importance Shift: SHAP Evidence

While Figure 3 quantifies the magnitude of leakage inflation, Figure 4 reveals its mechanism by comparing mean absolute SHAP values for the four tree-based ensemble models under Conditions A and B. Under Condition A, `survival_months` (red bar) clearly dominates in all four models.

Across all four models under Condition A, `survival_months` (red bar) clearly dominates the importance ranking, filling the entire panel width in

each subplot. In Gradient Boosting, its mean absolute SHAP value is 1.02 versus 0.30 for `node_positive_ratio` (second place), a dominance ratio of about 3.4:1. In XGBoost, the values are 1.84 versus 0.62, a ratio of about 3.0:1. The scale difference between the XGBoost and Gradient Boosting panels because SHAP magnitudes are model-specific, but the dominance pattern itself, survival_months overwhelming every clinical predictor, is consistent across both models.

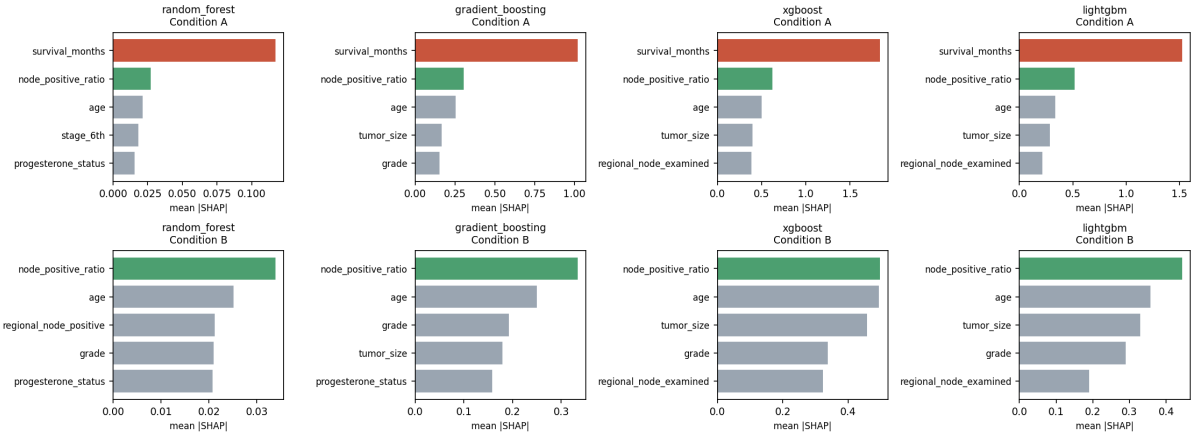


Figure 4. Feature importance shift from Condition A to Condition B

In the Random Forest panel, even with smaller absolute SHAP values, the `survival_months` bar (≈0.115) is still four to five times larger than the next feature (≈0.025). The other 17 clinical variables together contribute less predictive weight than `survival_months` alone. This pattern shows that Condition A models have not learned prognosis but a retrospective proxy for survival time. A model with implicit access to how long a patient lived is not predicting death.

This is a tabular instance of *shortcut learning*; when a statistically dominant but clinically non-actionable feature is available, models default to it rather than to the harder, generalizable signal in the remaining covariates. Once that shortcut is removed, the feature that fills the gap in which nodal involvement is a primary determinant of AJCC staging and long-term prognosis; the SHAP shift therefore reflects the model being forced back onto genuinely prognostic information rather than an arbitrary substitution.

3.4 Fixed-Horizon Evaluation: Condition C

Condition C offers a third independent test of the leakage-free performance ceiling and adds a clinically well-defined prediction target. As shown in Figure 2, the fixed-horizon reformulation changes both the effective cohort size (3,956 for H = 36m; 3,270 for H = 60m) and the event rate, most notably at H = 36 months, where only 5.94% of the cohort experienced the event. This makes C-36m the most imbalance-sensitive condition.

For H = 36 months, the best ROC-AUC is 0.767; for H = 60 months, it is 0.740. Both lie within the ~0.74–0.77 range found under Condition B, and the pairwise bootstrap intervals for Gradient Boosting under B [0.689, 0.785], C-36m [0.675, 0.819], and C-60m [0.674, 0.799] all overlap. This three-way agreement strongly indicates that ~0.74–0.77 AUC is a true empirical ceiling for this dataset and feature set.

Condition A's best AUC of 0.889 clearly shows the leakage effect. Leakage inflates performance by roughly 0.12–0.15 AUC points. A researcher selecting models based on Condition A would effectively rank them by how well they learn a disguised outcome, not by their true prognostic value.

3.5 Calibration and the Discrimination-Calibration Trade-off

The main contribution of this study is characterizing probability calibration across imbalance-handling variants, using the Brier Skill Score (BSS) to enable fair comparison across conditions with different event rates. This calibration analysis has not previously been reported for this dataset.

Figure 5 shows how sharply calibration differs across conditions. In Condition A (top-left), all four curves span the full [0, 1] range, with several models concentrating predictions near 0 and 1. This reflects the presence of `survival_months`, which gives a strong signal: patients surviving many months receive near-zero death probabilities, and those dying early receive near-one.

Condition A curves show some systematic overconfidence but roughly follow the diagonal, producing positive BSS values. For Gradient Boosting, BSS = 0.434 in Condition A—the best calibration in the study.

In Condition B, the curves compress toward the left of the x-axis, concentrating between 0 and ~0.5, as models estimate a 15.3% base-rate outcome without a retrospective time proxy. Gradient Boosting's BSS drops to 0.118, and Logistic Regression (L1) without resampling achieves BSS = 0.1255, the best calibration

among leakage-free models. In Condition C-60m, the curves show a similar compression, with Gradient Boosting's BSS improving to 0.146, consistent with its ~14% event rate being close to the Condition B event rate.

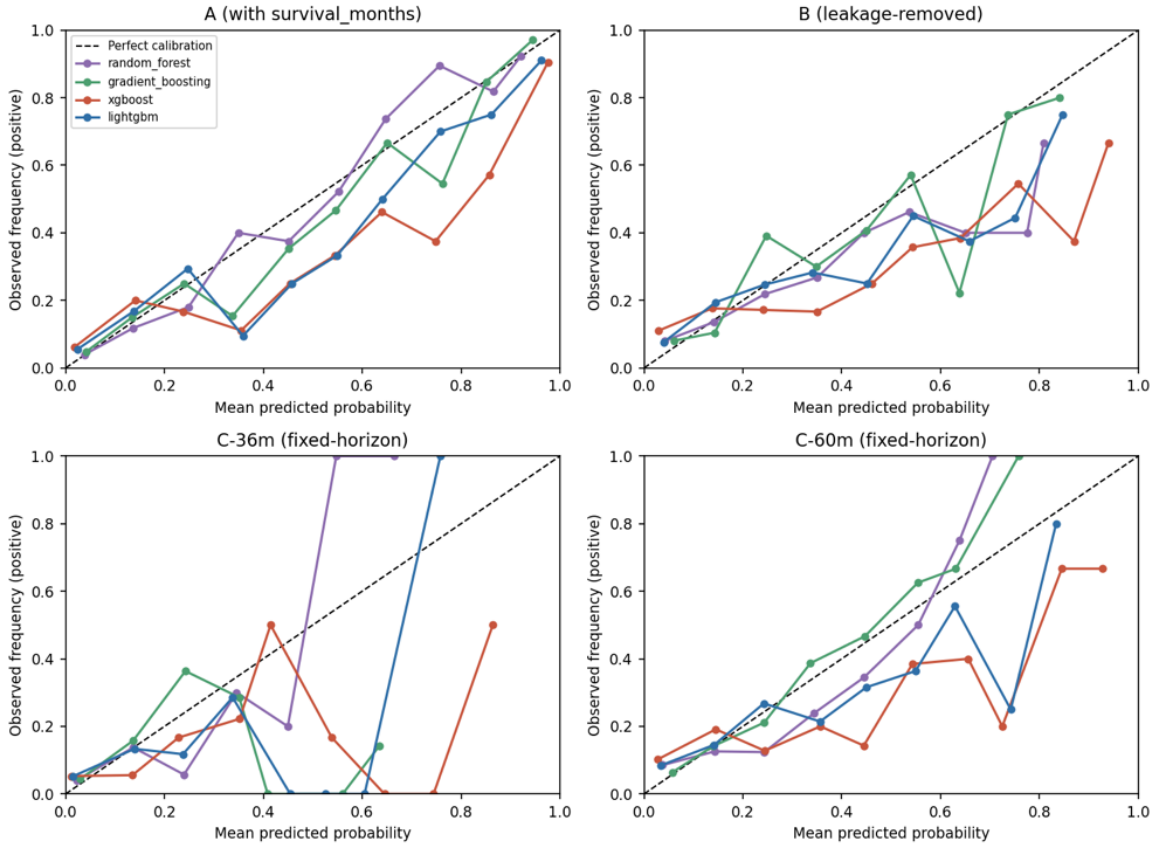


Figure 5. Reliability diagrams for four tree-based models

The Condition C-36m panel is the most distinctive: the curves collapse near the origin, with the x-axis only reaching about 0.5–0.6. Bins above this are empty, indicating a model that correctly identifies the low event rate (5.94%) but cannot meaningfully distinguish higher-risk patients within this compressed probability range.

Gradient Boosting's BSS in C-36m is about 0.001—not evidence of miscalibration, but a result of the near-zero Brier baseline that leaves little room to improve. The Decision Tree's BSS of −0.203 in Condition A (the only negative value there) shows that single-tree leaf-node probability estimates are pathologically overconfident, with or without the leakage feature.

Figure 6 crystallizes the core trade-off in handling imbalance. The `none` points (blue) cluster near BSS = 0 in all panels, over a narrow F1(dead) range, representing models that neither greatly improve minority detection nor severely harm calibration.

In contrast, `class_weight` (orange) and SMOTE (purple) points shift right and down: F1(dead) rises by up to 10–20 percentage points, while BSS plunges deep into negative values. The worst case appears in the C-36m panel, where AdaBoost with class weighting reaches BSS ≈ −2.578, meaning its probabilities are 3.6 times less accurate than simply predicting the 5.94% base rate for all patients.

This collapse has a clear statistical explanation: SMOTE and class weighting alter the effective class balance the model is optimized against during training, shifting the learned decision function's implicit prior away from the true test-time prevalence; the resulting scores still rank patients sensibly, which is why F1(dead) improves, but their absolute magnitudes no longer track observed event rates, which is why calibration degrades even as discrimination need not suffer.

The discrimination–calibration trade-off is highlighted by comparing the best-F1(dead) and best-BSS models in each clean condition. In Condition B, the F1(dead)-maximizing model is Logistic Regression (L2) with BorderlineSMOTE (F1(dead) = 0.414, BSS = −0.441), while the BSS-maximizing model is Logistic Regression (L1) without resampling (BSS = 0.1255). In Condition C-36m, Gaussian Naïve Bayes without resampling achieves the highest F1(dead) = 0.309, with BSS = −1.078. In Condition C-60m, Gaussian Naïve Bayes with BorderlineSMOTE attains F1(dead) = 0.420 but is poorly calibrated (BSS = −0.579). Across all three

clean conditions, Logistic Regression (L1) without resampling yields the best calibration: BSS = 0.1255 (B), 0.0942 (C-36m), and 0.1488 (C-60m). Thus, the simplest model with minimal intervention is the most reliable probability estimator.

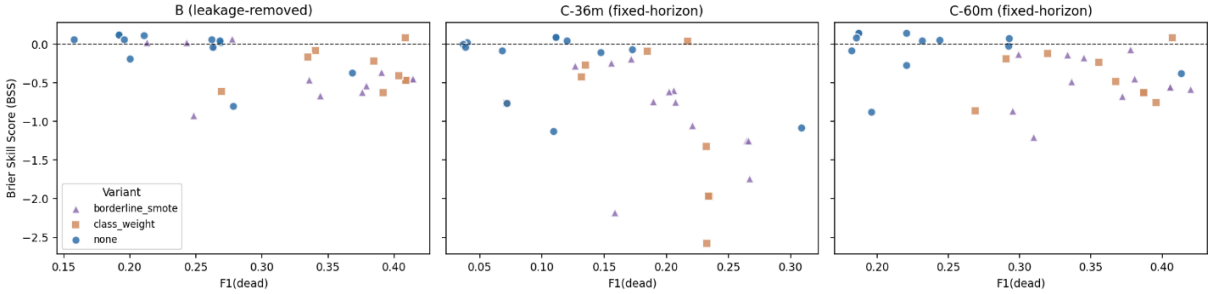


Figure 6. Discrimination-calibration trade-off

For clinical use, this means that models tuned for minority-class recall with SMOTE or class weighting are likely to be poorly calibrated. When predictions are used for absolute risk communication or to set treatment thresholds BSS and reliability diagrams should be reported alongside F1(dead).

Figure 6 suggests these goals may not be simultaneously achievable with current imbalance-handling methods on this dataset.

3.6 Statistical Validation and Model Ranking

The DeLong and McNemar paired tests show that the performance differences reported above are statistically robust rather than due to test-set sampling noise. Table 1 summarizes all pairwise comparisons among the four headline models—Decision Tree (DT), Gradient Boosting (GBM), Random Forest (RF), and Logistic Regression (LR)—reporting ROC-AUC values and DeLong p-values, as well as McNemar discordant-pair counts (b / c) and p-values for each condition.

Table 1. Pairwise statistical comparisons of model discrimination

Condition	Comparison	AUC (A)	AUC (B)	ΔAUC	DeLong p	McNemar (b / c)	McNemar p	Significance
A	GBM vs RF	0.889	0.877	+0.012	0.1927	9 / 6	0.6056	ns
A	GBM vs LR	0.889	0.875	+0.014	0.0572	23 / 16	0.3367	ns
A	RF vs LR	0.877	0.875	+0.002	0.8199	19 / 15	0.6069	ns
B	GBM vs RF	0.737	0.712	+0.025	0.1171	13 / 5	0.0990	ns
B	GBM vs LR	0.737	0.742	−0.005	0.7441	10 / 12	0.8312	ns
B	RF vs LR	0.712	0.742	−0.030	0.1118	8 / 18	0.0776	ns
C-36m	GBM vs RF	0.750	0.707	+0.043	0.0634	0 / 9	0.0077	McNemar only
C-36m	GBM vs LR	0.750	0.767	−0.017	0.2318	0 / 5	0.0736	ns
C-36m	RF vs LR	0.707	0.767	−0.060	0.0158	5 / 1	0.2207	DeLong only
C-60m	GBM vs RF	0.736	0.676	+0.060	0.0024	12 / 10	0.8312	DeLong only
C-60m	GBM vs LR	0.736	0.740	−0.004	0.8196	7 / 5	0.7728	ns
C-60m	RF vs LR	0.676	0.740	−0.064	0.0041	10 / 10	0.8231	DeLong only

DT = Decision Tree; GBM = Gradient Boosting; RF = Random Forest; LR = Logistic Regression (L2). ΔAUC = AUC(A) − AUC(B). b / c = McNemar discordant pairs (model A correct & B wrong / A wrong & B correct). Significance flag: "both" = both tests p < 0.05; "DeLong only" / "McNemar only" = exactly one test p < 0.05; "ns" = neither. Comparisons against the Decision Tree baseline is reported in Supplementary Table S1.

The Decision Tree performs far worse than all three stronger model families in every condition. In all 12 DT-versus-other comparisons (three pairs × four conditions), both tests reject the null hypothesis (all DeLong and McNemar p < 0.0001). This gap already appears in Condition A, where the Decision Tree's AUC is 0.717 versus 0.889 for Gradient Boosting and the tight 0.875–0.889 range for GBM, RF, and LR.

Across the three stronger families, no pairwise AUC differences are statistically significant in Conditions A or B (0 of 3 DeLong pairs in each). Thus, with any reasonably capable learner, algorithm choice does not meaningfully affect discrimination under leakage-free evaluation. Large model-to-model gaps reported in prior work on this dataset are therefore more likely due to target leakage and unstratified evaluation than to genuine superiority of any single method.

Under fixed-horizon targets, the tie breaks in a model-specific way that identifies Random Forest (RF) as the weak link. DeLong tests show RF is significantly worse than Logistic Regression (LR) in C-36m (p = 0.0158) and worse than both Gradient Boosting (GBM, p = 0.0024) and LR (p = 0.0041) in C-60m, while GBM and LR remain statistically indistinguishable at both horizons. A plausible explanation is the severe class imbalance (5.94% event rate at 36 months): RF's probability averaging pulls minority-class scores toward the majority, degrading ranking performance relative to boosting and regularized linear models.

A further observation concerns the relationship between the two tests, with disagreements concentrated in the imbalanced fixed-horizon conditions. In C-60m, for instance, GBM and RF differ significantly in AUC (DeLong p = 0.0024) but not in hard-label error rate

(McNemar p = 0.83); the opposite holds in C-36m (DeLong p = 0.063, McNemar p = 0.0077).

DeLong compares full probability rankings, while McNemar only compares binary decisions at the 0.5 threshold. With strong class imbalance, hard predictions collapse to the majority class, so models can have very different rankings yet nearly identical 0.5-threshold labels, and vice versa.

Overall, these results show that Logistic Regression performs as well as or better than all more complex alternatives. It is statistically indistinguishable from GBM and Random Forest in Conditions A and B and significantly outperforms Random Forest in both fixed-horizon conditions.

This reflects a common finding in clinical prediction: well-regularized linear models often match the discrimination of flexible ensembles in moderate-dimensional, small-to-medium-sample settings while providing much better interpretability and calibration.

4. Conclusions

This study set out to quantify how much of the 95–98% accuracy reported in 2022–2025 SEER breast cancer studies reflect genuine predictive performance rather than methodological artifact — the objective set out in the Introduction. Very little of it does: the reported accuracy and AUC figures trace to the same root cause, using Survival Months as an input feature when predicting the binary outcome Status, a direct form of target leakage in which models learn to encode a retrospective survival duration rather than perform prospective prognosis.

Once this leakage is removed and evaluation is corrected for class imbalance and censoring, discrimination settles at a substantially lower, but genuine and reproducible, ceiling of 0.74–0.77 AUC the realistic benchmark this audit set out to establish. Beyond this benchmark, the practical implications for registry-based clinical prediction research are fourfold.

Together, these findings lead to four actionable recommendations. First, exclude `Survival Months`—and any retrospective post-diagnosis variable—from feature sets for cross-sectional or prospective classifiers. Second, accuracy alone is an inadequate primary metric on imbalanced registry data; complement it with F1(dead), ROC-AUC, and PR-AUC for the minority class.

Third, fixed-horizon binary reformulation with proper censoring exclusion is a more clinically defensible target than the cross-sectional `Status` label and yields performance estimates that closely match the leakage-free cross-sectional benchmark. Fourth, imbalance-handling techniques that improve recall-based metrics should be vetted against calibration measures before deployment in any setting where probability magnitudes matter.

This study has several limitations. The analysis relies on a single, moderately sized registry extract (n = 4,023). The 0.74–0.77 AUC ceiling may reflect limits of available clinical features or cohort-specific artifacts and requires validation in larger, multi-institutional datasets. We did not perform systematic hyperparameter optimization to avoid tuning confounders in the A/B/C comparison, so individual models may not represent each algorithm's best performance with this feature set.

While this audit is grounded in structured registry data, the two failure modes identified here are not specific to tabular models. In image-based and multimodal deep learning for oncology, Survival Months-style leakage rarely appears as an explicit input column, as CNNs trained on chest radiographs have been shown to exploit hospital- and scanner-specific artifacts rather than pathology.

In multimodal architectures that fuse imaging with structured clinical variables, the same mechanism quantified here for Survival Months can re-enter through the tabular branch even when the imaging branch is leakage-free, so leakage audits of fused models should isolate each modality's contribution separately.

Future work should evaluate full survival models (e.g., Cox proportional hazards, random survival forests) that preserve continuous time-to-event information instead of binarizing horizons, and test whether calibration collapse under imbalance handling can be reduced with post-hoc methods such as Platt scaling or isotonic regression.

Declarations

AI Usage Statement

The authors used Claude solely to improve the readability and language of the manuscript; all scientific content, analysis, and conclusions are the authors' own.

Author Contributions

S.N.R: conceptualization, writing – original draft; V.M.N: data curation, formal analysis; E.A.J: methodology; V.B.L: review; E.M: editing.

Funding Statement

This research was funded by the Lembaga Layanan Pendidikan Tinggi Wilayah III (LLDIKTI III) Jakarta, Indonesia, through Reputable Journal Publication Assistance Program. The authors express their sincere gratitude for the support provided.

Conflict of Interest Statement

The authors declare no conflict of interest.

Data and Code Availability

The processed dataset and the analysis code supporting the findings of this study are not publicly available due to institutional policy. Data and code are available from

the corresponding author upon reasonable request.

References

[1] E. T. Sedeta, B. Jobre, and B. Avezbakiyev, "Breast cancer: Global patterns of incidence, mortality, and trends.," *J. Clin. Oncol.*, vol. 41, no. 16_suppl, pp. 10528–10528, Jun. 2023, doi: 10.1200/JCO.2023.41.16_suppl.10528.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] H. Gu *et al.*, "Global burden and trends of breast cancer: GLOBOCAN 2022 estimates of incidence and mortality in 185 countries," *Chin. Med. J. (Engl.)*, vol. 139, no. 3, pp. 404–414, Feb. 2026, doi: 10.1097/CM9.0000000000003921.

[4] C. Chu, N. White, and P. Stone, "Prognostication in palliative care," *Clin. Med.*, vol. 19, no. 4, pp. 306–310, Jul. 2019, doi: 10.7861/clinmedicine.19-4-306.

[5] M. Maltoni *et al.*, "Prognostication in palliative radiotherapy—ProPaRT: Accuracy of prognostic scores," *Front. Oncol.*, vol. 12, p. 918414, Aug. 2022, doi: 10.3389/fonc.2022.918414.

[6] J. TENG, "SEER Breast Cancer Data." IEEE DataPort, Jan. 18, 2019. doi: 10.21227/A9QY-PH35.

[7] J. Li *et al.*, "Predicting breast cancer 5-year survival using machine learning: A systematic review," *PLOS ONE*, vol. 16, no. 4, p. e0250370, Apr. 2021, doi: 10.1371/journal.pone.0250370.

[8] S. Kagan *et al.*, "Survival prediction in metastatic breast cancer using artificial intelligence: a scoping review," *Curr. Opin. Support. Palliat. Care*, vol. 19, no. 4, pp. 279–289, Dec. 2025, doi: 10.1097/SPC.0000000000000780.

[9] P. Manikandan, U. Durga, and C. Ponnuraja, "An integrative machine learning framework for classifying SEER breast cancer," *Sci. Rep.*, vol. 13, no. 1, p. 5362, Apr. 2023, doi: 10.1038/s41598-023-32029-1.

[10] R. Watrianthos, R. Rayendra, E. Asri, Y. Yuhefizar, and H. Humaira, "Threshold-Optimized and Calibrated Logistic Regression for Breast Cancer Classification," *Salud Cienc. Tecnol.*, vol. 5, p. 2241, Oct. 2025, doi: 10.56294/saludcyt20252241.

[11] A. M. Alharthi, M. H. Lee, and Z. Y. Algama, "Improving the Diagnosis of Breast Cancer Using Regularized Logistic Regression with Adaptive Elastic Net," *Univers. J. Public Health*, vol. 9, no. 5, pp. 317–323, Oct. 2021, doi: 10.13189/ujph.2021.090514.

[12] H. X. Liu *et al.*, "Diagnosing Breast Cancer Based on Support Vector Machines," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 3, pp. 900–907, May 2003, doi: 10.1021/ci0256438.

[13] M. Kavitha, G. Lavanya, J. Janani, and Balaji. J, "ENHANCED SVM CLASSIFIER FOR BREAST CANCER DIAGNOSIS," *Int. J. Eng. Technol. Manag. Res.*, vol. 5, no. 3, pp. 67–74, Feb. 2020, doi: 10.29121/ijetmr.v5.i3.2018.178.

[14] A. Rafiq *et al.*, "Detection and Classification of Histopathological Breast Images Using a Fusion of CNN Frameworks," *Diagnostics*, vol. 13, no. 10, p. 1700, May 2023, doi: 10.3390/diagnostics13101700.

[15] P. Ghosh and D. Chatterjee, "Comparative Analysis of Machine Learning Algorithms for Breast Cancer Classification: SVM Outperforms XGBoost, CNN, RNN, and Others," Apr. 26, 2024, *Bioinformatics*. doi: 10.1101/2024.04.22.590658.

[16] M. Chetry *et al.*, "Early detection and analysis of accurate breast cancer for improved diagnosis using deep supervised learning for enhanced patient outcomes," *PeerJ Comput. Sci.*, vol. 11, p. e2784, Apr. 2025, doi: 10.7717/peerj-cs.2784.

[17] Agus Perdana Windarto, Anjar Wanto, S Solikhun, and Ronal Watrianthos, "A Comprehensive Bibliometric Analysis of Deep Learning Techniques for Breast Cancer Segmentation: Trends and Topic Exploration (2019-2023)," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 7, no. 5, pp. 1155–1164, Oct. 2023, doi: 10.29207/resti.v7i5.5274.

[18] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[19] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nat. Commun.*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

[20] L. Sasse *et al.*, "Overview of leakage scenarios in supervised machine learning," *J. Big Data*, vol. 12, no. 1, p. 135, May 2025, doi: 10.1186/s40537-025-01193-8.

[21] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit. Health*, vol. 2, no. 11, p. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290.

[22] T. G. Baidoo and H. Rodrigo, "Data-driven survival modeling for breast cancer prognostics: A comparative study with machine learning and traditional survival modeling methods," *PLOS One*, vol. 20, no. 4, p. e0318167, Apr. 2025, doi: 10.1371/journal.pone.0318167.

[23] S. Kaindal and B. Venkataramana, "A comparative analysis of parametric survival models and machine learning methods in breast cancer prognosis," *Sci. Rep.*, vol. 15, no. 1, p. 31288, Aug. 2025, doi: 10.1038/s41598-025-15696-0.

[24] M. Cruz-Fernandez *et al.*, "Predicting Breast Cancer Mortality Using SEER Data: A Comparative Analysis of L1-Logistic Regression and Neural Networks," *Technologies*, vol. 14, no. 1, p. 66, Jan. 2026, doi: 10.3390/technologies14010066.

[25] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innov.*, vol. 3, no. 5, p. e136, Oct. 2024, doi: 10.1002/cai2.136.

[26] Y. Liu, Y. Fu, Y. Peng, and J. Ming, "Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory," *Heliyon*, vol. 10, no. 2, p. e24876, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24876.

[27] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, *arXiv*. doi: 10.48550/ARXIV.1705.07874.