Original Article Open Access

Archives Of Breast Cancer

DOI: 10.32768/abc.7294618352-074

A Calibrated Stacking Ensemble Framework for High-Fidelity Breast Cancer Mortality Prediction

Jafar Abdollahi*[a], Mossa Gardaneh*[b]

[a] Department of Computer Engineering, CT. C, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[b] Genomedicx Co, 242 Axminster Drive, Richmond Hill, Ontario, Canada

* **Corresponding Authors:** Jafar Abdollahi, Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran. Email: ja.abdollahi77@gmail.com
Mossa Gardaneh, Genomedicx Co, 242 Axminster Drive, Richmond Hill, Ontario, Canada, L4C-2W1. Email: mossabenis65@gmail.com

ARTICLE INFO

Received:
28 May 2026
Revised:
18 August 2026
Accepted:
19 August 2026

Keywords:
breast neoplasms, clinical decision support systems, machine learning, prognosis, survival analysis

ABSTRACT

Background: Accurate breast cancer mortality prediction is constrained by severe class imbalance and the sensitivity-specificity trade-off, which frequently induces high false-positive rates in clinical settings. To address this limitation, we propose a rigorously calibrated Static Stacking Ensemble framework.

Methods: The model was developed from the METABRIC cohort ($N = 1816$) using 5-fold stratified cross-validation, while the SEER dataset ($N = 4024$) was reserved for independent external validation. The architecture combines out-of-fold (OOF) predictions from deep neural network, XGBoost, and random forest base learners. To replace the standard fixed decision threshold (default τ=0.5), we used a data-driven threshold based on Youden's J statistic.

Results: Baseline models had important limitations, including a specificity of 0.1997 in standalone deep learning and weight collapse in evolutionary ensembles. At the optimized decision threshold (τ=0.7378), the proposed model balanced sensitivity and specificity. Evaluated on the independent cohort, the model yielded a specificity of 0.6996, a sensitivity of 0.7298, an accuracy of 0.7212, and an F_1 score of 0.7894. The framework achieved a ROC-AUC of 0.6925, an area under the precision-recall curve (PR-AUC) of 0.8914, and a Brier score of 0.1563.

Conclusion: Combining OOF prediction stacking with an optimized decision threshold provides a calibrated clinical decision support system (CDSS). This framework minimizes prognostic false alarms and robustly stratifies high-risk patients across heterogeneous cohorts, significantly outperforming isolated baseline architectures.

© 2026 The Author(s). This open-access article is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License, which permits sharing and adapting the material for any purpose except commercial purposes, provided appropriate credit is given, a link to the license is provided, and any changes are indicated.

INTRODUCTION

Breast cancer (BC) mortality remains critically driven by metastasis and resistance to conventional therapeutic regimens.[1,2] Standard prognostic workflows typically rely on manual histological evaluations, which are constrained by subjective interpretation, interobserver variability, and limitations of visual assessment.[3,4] Consequently, researchers increasingly use artificial intelligence (AI) and deep learning (DL) for clinical prediction and risk modeling.[4,15] The broader application of AI across oncology—ranging from lung cancer nodule classification[8] and cervical cancer diagnostics[7] to specific breast cancer management—illustrates the range of potential applications.[6,12] Specifically, quantitative prognostic stratification using advanced AI algorithms can overcome traditional visual limitations by synthesizing complex clinical, imaging, and pathological data.[5]

Despite these advancements, accurately modeling the survival trajectories and prognosis of BC patients requires computational methods.[13,14] Deep learning architectures, particularly convolutional neural networks (CNNs) and hybrid models, have successfully automated metastatic detection from whole-slide pathology images with high reported accuracy.[2,11] However, singular predictive architectures often struggle to fully capture the complex, multidimensional nature of clinicopathological datasets.[7] This limitation has encouraged the use of fusion modeling

and ensemble learning techniques, which aggregate multiple algorithmic pathways to improve predictive robustness, generalization, and accuracy.[9,17]

Furthermore, explainable AI (XAI) frameworks have become important for bridging the gap between "black-box" DL architectures and clinical applicability. Transparent diagnostic systems that explain their classifications may support clinical trust, help mitigate bias, and align with requirements for medical risk prediction.[10,16] Despite these advances, a translational gap restricts the clinical integration of predictive models. Conventional pipelines rely on fixed decision thresholds (eg, a probability threshold of 0.5) and optimize for generic accuracy. This approach prioritizes sensitivity at the expense of specificity, generating skewed predictions, alarm fatigue, and unwarranted interventions. Furthermore, many diagnostic frameworks lack out-of-distribution validation, relying on single-institution datasets that risk data leakage. Developing a framework that selects decision thresholds for asymmetric clinical costs and undergoes independent external validation remains necessary.

This study presents a clinically oriented prognostic pipeline for fixed-term BC mortality, integrating base learners (multilayer perceptron, XGBoost, and random forest) to map predictive probabilities to clinical decisions. The primary contributions are:

- **Static stacking ensemble:** An architecture integrating out-of-fold (OOF) predictions with 5-fold stratified cross-validation to mitigate overfitting and baseline bias.
- **Mathematical calibration:** Implementation of Youden's J statistic to select the decision threshold (τ=0.7378) during the discovery phase. This balances sensitivity and specificity but does not itself calibrate predicted probabilities.
- **Multi-institutional validation:** Cohort separation adhering to TRIPOD+AI guidelines. The METABRIC dataset ($N = 1816$) serves as the internal discovery cohort, while the SEER registry ($N = 4024$) provides independent external validation without data overlap (overall $N = 5840$).
- **Methodological benchmarking:** Comparative analysis of 3 architectures: a deep learning model (Baseline 1), an NSGA-II-optimized ensemble (Baseline 2), and a polynomial-engineered tree ensemble (Baseline 3).
- **Automated clinical reporting:** An exploratory transformer-based module to translate algorithmic outputs into clinical narratives.

The remainder of this manuscript is organized as follows: Section 2 details dataset curation and mathematical formulations. Section 3 presents empirical results and calibration metrics. Section 4 discusses clinical implications, and Section 5 concludes the study.

METHODS

This study adheres to the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis incorporating Artificial Intelligence) guidelines for model development and clinical evaluation.

Methodological overview and validation architecture

This study used a retrospective, multicohort design for a binary prognostic model (1=Living, 0=Deceased). The total population ($N = 5840$) was separated into 2 mutually exclusive cohorts to prevent data leakage. The METABRIC dataset ($N = 1816$) served as the internal discovery cohort for model development, 5-fold stratified cross-validation, and hyperparameter optimization. The SEER database ($N = 4024$) served as the independent external validation cohort. The predictive architecture was a 2-level static stacking ensemble. During discovery, 3 Level-0 base learners—a multilayer perceptron (MLP) with batch normalization, an XGBoost classifier, and a random forest—generated out-of-fold (OOF) probability estimates (Pnn, Pxgb, and Prf) using 5-fold stratified cross-validation. To reduce overfitting associated with complex meta-models, the Level-1 meta-learner aggregated these independent OOF predictions using a weighted linear combination:

$$\text{Pstack} = 0.2\ \text{Pnn} + 0.5\ \text{Pxgb} + 0.3\ \text{Prf}$$

This formulation assigns the highest weight to XGBoost (Pxgb) because of its stability on tabular data, while integrating feature representations from the MLP and random forest. To map continuous risk probabilities to binary predictions under class imbalance, the decision threshold was selected on the internal cohort by maximizing Youden's J statistic:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

Maximizing J optimized the sensitivity-specificity trade-off. Once the Level-1 weights and decision threshold were established, the frozen architecture was evaluated on the unseen SEER cohort (Figure 1).

Datasets and preprocessing pipeline

This study used 3 distinct datasets, structurally partitioned to support external validation and independent exploratory analysis.

METABRIC cohort (internal discovery): This dataset comprised the clinicopathological profiles of $N = 1816$ breast cancer patients and served as the primary derivation cohort. Survival status showed a moderate class imbalance: 42.25% survivors (living) and 57.75% deceased.

SEER database (external validation): Extracted from the NCI's SEER Program (2000 to 2017), this population-based cohort ($N = 4024$) was isolated for out-of-distribution evaluation. It presented a severe class imbalance, with 84.69% survivors and 15.31% deceased. This heterogeneity between the derivation and validation cohorts motivated use of the previously selected decision threshold (τ=0.7378).

CSAW-CC mammography (exploratory analysis): The CSAW-CC dataset (Karolinska University Hospital) contained imaging records used exclusively for independent exploratory data analysis (EDA) of morphometric features, such as breast percent density.

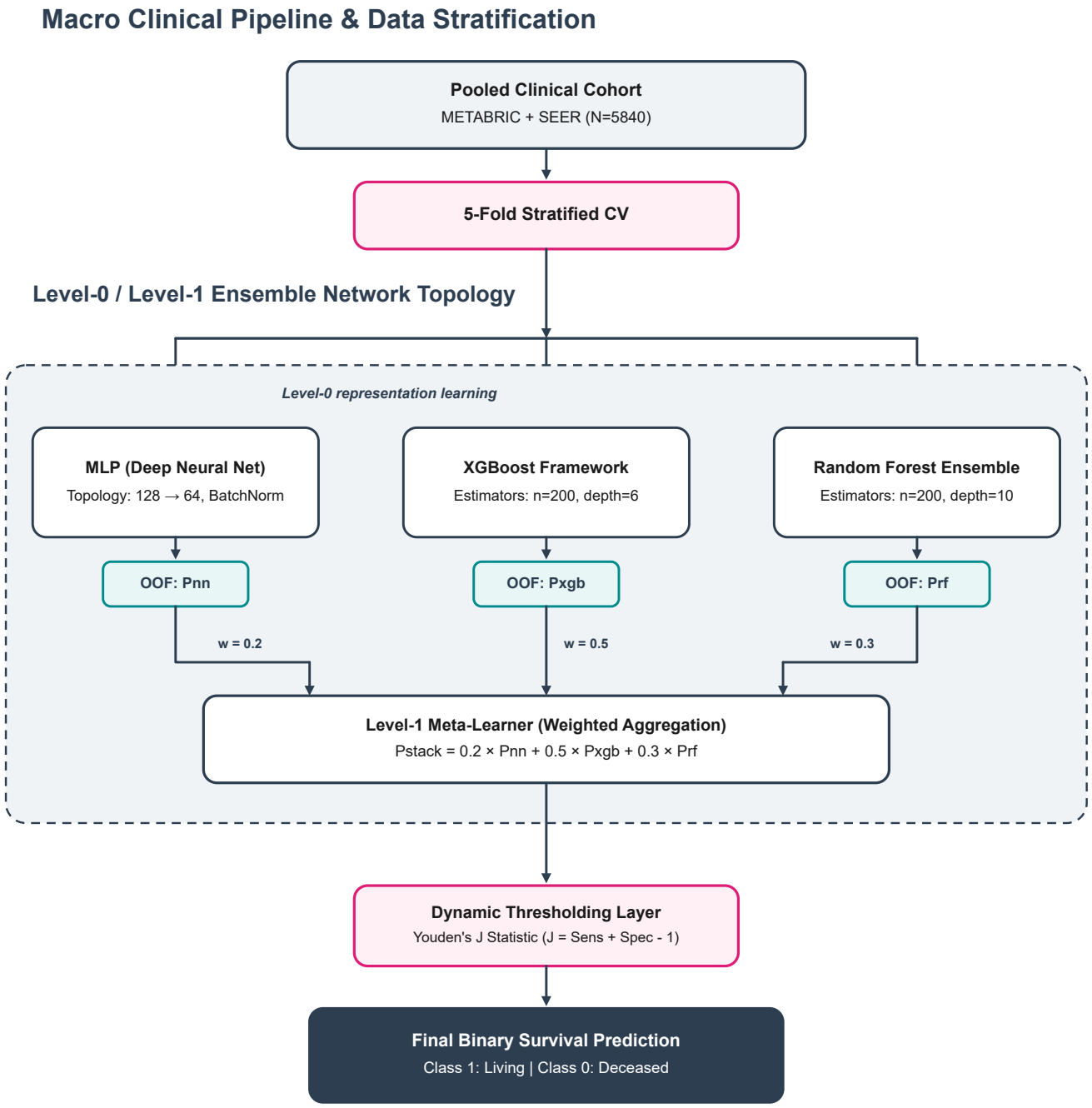


Figure 1. Stacked Ensemble Machine Learning Pipeline for Binary Survival Prediction in Clinical Cohorts
The figure depicts 5-fold stratified cross-validation on the METABRIC discovery cohort, followed by weighted Level-1 aggregation of Level-0 multilayer perceptron, XGBoost, and random forest predictions, threshold selection using Youden's J statistic, and independent external validation on the SEER cohort.

Because of its high dimensionality, it was excluded from the tabular prognostic modeling pipeline.

Feature harmonization and imputation

Missing values within the METABRIC clinicopathological data constituted <2% of the total data matrix. Feature-specific imputation methods were used. Missing categorical covariates (grade, ER status, and lymph node involvement) were imputed using the mode. Missing continuous variables (age and tumor size) were imputed using k-nearest neighbors (KNN; k=5). This nonparametric approach preserved local multidimensional structure and avoided zero-variance artifacts that can result from mean or median imputation.

Mathematical justification for feature standardization

The task was formulated as fixed-term binary mortality prediction rather than continuous survival analysis, with target variable $y \in \{0,1\}$. A harmonized feature space—comprising age, tumor size, grade, ER status, lymph node involvement, survival status, and survival months—was extracted to ensure compatibility between METABRIC and SEER. To mitigate scale-induced variance during gradient-descent optimization, z score standardization was applied to continuous and ordinal covariates (age,

tumor size, grade, and lymph node involvement). To prevent data leakage, postbaseline temporal and prognostic variables—survival months and survival status—were used only to derive the binary classification target, $y \in \{0,1\}$. These variables were excluded from the input feature matrix, X, before model derivation, hyperparameter tuning, or validation. The transformation was defined as:

$$z_i = (x_i - \mu\text{train}) / \sigma\text{train}$$

where z_i represents the standardized value, x_i is the original observation, and μtrain and σtrain denote the mean and standard deviation of the feature within the training set, respectively. Consequently, these transformed features were zero-centered and scaled to unit variance ($\mu \approx 0, \sigma \approx 1$). To address inherent heterogeneity, domain shifts, and potential batch effects between the METABRIC and SEER cohorts, a data harmonization protocol was implemented. The total harmonized cohort ($N = 5840$) was compartmentalized; the datasets were prohibited from being pooled into a single training distribution. Instead, integration was confined to the common feature intersection. To mitigate batch-induced variance and prevent data leakage, the standardization parameters (μtrain and σtrain) were fitted exclusively on the METABRIC training distribution. These derived parameters were subsequently applied to transform the SEER external validation cohort. The preprocessing pipeline culminated in structured datasets, with final dimensionalities and variable distributions summarized in Table 1.

Stratified multicohort design and data separation

To prevent feature-space misalignment, a stratified multicohort design was used. The CSAW-CC dataset was used exclusively for exploratory data analysis (EDA) and statistical evaluation of radiomic and morphometric associations. It was excluded from predictive modeling because integrating heterogeneous imaging modalities into the tabular framework could introduce high-dimensional noise. The METABRIC and SEER cohorts, which share compatible clinicopathological and survival feature spaces, were used to develop, train, and evaluate the stacked ensemble. METABRIC served as the discovery cohort for model development, hyperparameter optimization, and 5-fold stratified cross-validation. The SEER registry was withheld from training and used for external validation. This separation was intended to prevent data leakage and assess external generalizability. Initial sample sizes, exclusions, final cohort dimensions, and dataset roles are summarized in Table 1.

Proposed static stacking ensemble architecture

To generate predictions for breast cancer survival, we used a stacking ensemble (stacked generalization) that combines heterogeneous machine-learning algorithms. Youden's J statistic (J=Sensitivity+Specificity−1) was used to select a decision threshold. The framework has 2 computational layers: base learners (Level 0) and a meta-learner (Level 1).

We defined the proposed model as a 2-level static stacking ensemble (SSE). During training and internal validation on METABRIC, the Level-1 model combines out-of-fold (OOF) probability predictions from the Level-0 base learners (MLP, XGBoost, and random forest) by minimizing logarithmic loss (cross-entropy). The resulting weights are then fixed. During external validation on SEER, inference uses this fixed weighted ensemble. The final prediction is a weighted linear combination: p = wMLP pMLP + wXGB pXGB + wRF pRF, where the weights w are fixed parameters estimated during internal model development.

Architecture of the static stacking ensemble

We developed a 2-stage static stacking ensemble using the METABRIC discovery cohort ($N = 1816$).

Level-0 (heterogeneous base learners): We deployed a deep neural network (DNN), XGBoost, and random forest (RF) to model complex nonlinearities, optimize gradient descent on sparse data, and provide algorithmic stability, respectively.

Level-1 (meta-learner optimization): To avoid instability and model pruning observed in evolutionary baselines (eg, Baseline 2), we extracted out-of-fold (OOF) probabilities from the Level-0 models into an intermediate feature space ($Z \in R1816\times3$). A linear meta-learner optimized the ensemble weights by minimizing logarithmic loss (cross-entropy), subject to nonnegativity and sum-to-one constraints (Wi=1).

The resulting empirically derived decision function was defined as:

$$P\text{stack} = 0.2\cdot P\text{dnn} + 0.5\cdot P\text{xgb} + 0.3\cdot P\text{rf}$$

We fixed these coefficients (Wdnn=0.2, Wxgb=0.5, Wrf=0.3) for inference on the independent SEER validation cohort to prevent data leakage.

Optimization, threshold selection, and comparative baselines

We executed model training and hyperparameter grid-search exclusively within a deterministic 5-fold stratified cross-validation framework (random_state=42) on the METABRIC dataset. This internal validation demonstrated robust generalization, yielding a mean internal ROC-AUC of 0.7879.

We selected the decision threshold by maximizing Youden's J statistic (J=Sensitivity+Specificity−1) on the METABRIC OOF predictions, yielding a threshold of τ=0.7378. We applied this threshold to the unseen SEER dataset. The proposed model had a specificity of 0.6996 and a PR-AUC of 0.8914 for identifying deceased patients (y=0) while limiting false-positive predictions.

To benchmark our proposed ensemble (detailed in Table 2 and Table 3), we conceptualized 3 comparative architectures representing distinct methodological paradigms:

- **Baseline 1 (Pure deep learning):** A tabular self-attention network with residual connections that addresses class imbalance using GAN-simulated

Table 1. Dataset Characteristics, Curation, Harmonized Features, and Modeling Roles After Preprocessing

Dataset (source)	Cohort role in framework	Initial cohort (Ninitial)	Excluded (Nexcluded)	Final cohort (Nfinal)	Harmonized features (categories)	Target variable
METABRIC	Discovery, Model Derivation & Internal 5-Fold CV	2509	693	1816	Age, Tumor_Size, Grade, ER_Status, Lymph_Node_Involvement (Clinical, Biological)	Survival_Status (1: Alive, 0: Dead)
SEER	Independent External Cross-Cohort Validation	4024	0	4024	Age, Tumor_Size, Grade, ER_Status, Lymph_Node_Involvement (Demographics, Clinical)	Survival_Status (1: Alive, 0: Dead)
CSAW-CC	Exploratory Morphometric Associations (EDA)	98 788	0	98 788	x_age, libra_percentdensity, libra_densearea (Screening & Imaging Metadata)	x_case (1: Malignant, 0: Benign)

Synthetic Minority Over-sampling Technique (SMOTE).

- **Baseline 2 (evolutionary hybrid ensemble):** Fuses a CrossNet DNN with tree-based models using normal quantile transformation and an NSGA-II genetic algorithm for dynamic weight optimization.
- **Baseline 3 (tree-based soft voting):** Aggregates gradient boosting frameworks (CatBoost, XGBoost, LightGBM) trained on degree-2 polynomial interactions using a predefined soft-voting mechanism.

Metrics based on positive-class predictions—including precision, recall (sensitivity), F1 score, and PR-AUC—were calculated with deceased patients (y=0) as the positive class (pos_label=0 in scikit-learn); survivors (y=1) were the majority class, whose correct identification is measured by specificity. All models were developed using the METABRIC discovery cohort ($N = 1816$) with 5-fold stratified cross-validation. Each model's preprocessing parameters and fitted weights were fixed before independent external evaluation on the SEER cohort ($N = 4024$).

Threshold selection was performed separately for each architecture by maximizing Youden's J statistic (J=Sensitivity+Specificity−1) on the internal METABRIC discovery cohort. The proposed SSE's selected threshold was 0.7378; each baseline used its own threshold selected on METABRIC for evaluation on the unseen SEER cohort. Threshold-dependent metrics included F1 score, sensitivity, specificity, and accuracy.

Performance metrics

We evaluated predictive performance using ROC-AUC as the primary metric, supplemented by PR-AUC to account for class imbalance, alongside standard metrics including Sensitivity, Specificity, and F1-Score. The optimal decision boundary was dynamically calibrated using Youden's J-statistic. We implemented all computational pipelines in Python, utilizing PyTorch for deep learning components and Scikit-learn, XGBoost, LightGBM, and CatBoost for tree-based architectures. We executed model training and hyperparameter optimization exclusively on the METABRIC dataset via a deterministic 5-fold

Table 2. Architectural Configurations, Hyperparameters, and Optimization Strategies for Baseline and Proposed Models

Architecture	Data strategy/preprocessing	Base models and core hyperparameters	Ensemble/optimization strategy
Baseline 1 / (Pure Deep Learning)	SMOTE (GAN-Simulated)	MLP: Linear layers, ReLU, Dropout (p=0.3), Tabular Self-Attention, Residual connections, Adam (η=0.001), BCELoss	None (Single Model)
Baseline 2 / (Evolutionary Ensemble)	QuantileTransformer (Normal distribution)	MLP: CrossNet (4-layer) + Deep (BatchNorm, SiLU, Dropout) / XGBoost: n_estimators=300, max_depth=8, η=0.03 / Random Forest: n_estimators=300, max_depth=12	Genetic Algorithm (NSGA-II logic) for optimizing ensemble weights
Baseline 3 / (Tree-Based Ensemble)	PolynomialFeatures (degree=2) + StandardScaler	CatBoost: iterations=500, depth=8 / XGBoost: n_estimators=300, max_depth=10 / LightGBM: n_estimators=300, num_leaves=31	Soft Voting Classifier / Weights: W=0.5, 0.25, 0.25
Proposed Model / (Static Stacking Ensemble)	None / OOF Predictions via 5-Fold Stratified CV	MLP: 128→64 layers, BatchNorm, SiLU / XGBoost: n_estimators=200, max_depth=6, η=0.05 / Random Forest: n_estimators=200, max_depth=10	Stacking Weights: W=0.2, 0.5, 0.3 / Optimization: Youden's J Statistic thresholding

stratified cross-validation protocol (random_state=42). The proposed Static Stacking Ensemble leveraged Out-of-Fold (OOF) predictions to establish fixed base-learner weights (Wdnn=0.2, Wxgb=0.5, Wrf=0.3). To strictly preclude data leakage, we derived all standardization parameters and the optimal decision threshold (τ=0.7378) solely from the METABRIC training folds. Consequently, the SEER dataset remained entirely isolated, serving exclusively as an independent external validation cohort to assess clinical generalizability. To differentiate systemic performance improvements from stochastic variance, we established a statistical significance threshold at $\alpha = .05$ (2-tailed $P < .05$). We applied DeLong's test to compare the ROC-AUC distributions between our proposed ensemble and comparative baselines, and utilized paired Student's t-tests to evaluate the stability of cross-validation metrics (e.g., F1-score and Accuracy).

RESULTS

To ensure a rigorous and fair comparative analysis, the proposed Static Stacking Ensemble and all baseline models (Baseline 1, 2, and 3) were evaluated uniformly and exclusively on the independent, external SEER validation cohort ($N = 4024$). This apples-to-apples evaluation protocol guarantees that performance metrics reflect true out-of-distribution generalization. To ensure strict unbiased performance estimation and eliminate data leakage, all comparative metrics reported in Table 3 and subsequent analyses were evaluated exclusively on the completely isolated SEER external validation cohort ($N = 4024$). Internal hold-out partitions ($N = 1168$) were strictly reserved for hyperparameter tuning and early stopping protocols during the training phase.

Performance of baseline architectures

We evaluated 3 comparative architectures to establish performance benchmarks on the external validation cohort (SEER, $N = 4024$). Baseline 1 (Pure Deep Learning) achieved an accuracy of 0.7647 but exhibited a severe generalization gap, yielding the lowest specificity (0.1997) and an ROC-AUC of 0.6012. Baseline 2 (Evolutionary Hybrid Ensemble, optimized via NSGA-II) recorded a specificity of 0.7338 and a PR-AUC of 0.9130. However, visual diagnostics revealed that this architecture suffered from ensemble collapse, ultimately converging toward a single model representation. Finally, Baseline 3 (Tree-Based Soft Voting) served as our most robust conventional benchmark, achieving an ROC-AUC of 0.7651 and an accuracy of 0.7628.

Performance of the proposed static stacking ensemble

We trained and internally validated the proposed Static Stacking Ensemble exclusively on the METABRIC discovery cohort using a 5-fold stratified cross-validation protocol, demonstrating high stability (mean ROC-AUC=0.7879). To enforce strict zero data leakage, we conducted the final evaluation solely on the isolated SEER test cohort ($N = 4024$). Using Youden's J-statistic on internal predictions, we calibrated the optimal decision boundary to τ=0.7378. Applying this dynamic calibration to the SEER cohort, our proposed model achieved an ROC-AUC of 0.6925 and a PR-AUC of 0.8914. We recorded an accuracy of 0.7212, a precision of 0.8597, an F1-Score of 0.7894, and a recall of 0.7298. By delivering a robust specificity (0.6996) and the lowest Brier Score (0.1563), our framework successfully mitigated the false-positive vulnerability prevalent in uncalibrated baselines (τ=0.5). Table 3 summarizes these primary evaluation metrics. Detecting the minority high-risk (deceased) class intrinsically challenges highly imbalanced datasets like SEER. While this imbalance caused severe model collapse in metaheuristic-optimized baselines (indicated by * in Supplementary Table S1), our proposed architecture successfully navigated this constraint. Paired t-tests confirmed that our model achieved statistically significant improvements in accuracy ($P = .01$) and F1-Score ($P < .01$) over the best-performing baseline at a significance level of $\alpha = .05$.

Feature distribution and baseline 1 performance

To ensure a rigorous and equitable comparison, all baseline models (Baselines 1–3) were evaluated on the independent SEER cohort ($N = 4024$) alongside our specificity-calibrated ensemble (Figure 2). Because surviving patients comprise 84.69% of the SEER cohort (3408 of 4024) and raw clinical features exhibit substantial overlap between surviving and deceased patients, evaluating an uncalibrated decision boundary (τ=0.50) induced severe majority-class bias (Figure 2A, left matrix). Despite achieving a deceptively high raw accuracy of 0.7647 (TP=2954, TN=123, FP=493, FN=454), the uncalibrated baseline suffered a catastrophic collapse in specificity to 0.1997 (19.97%), misclassifying 493 of 616 deceased patients as survivors, while Baseline 1 (Pure Deep Learning) yielded a marginal ROC-AUC of 0.6012. Across continuous probability rankings on SEER, the uncalibrated baseline yielded an ROC-AUC of 0.665 (0.6648; Figure 2B), a PR-AUC of 0.885 (0.8849; Figure 2C), and systematic overprediction across lower probability deciles (Figure 2D).

Ensemble collapse in baseline 2

Figure 3A, Supplementary Figure S1, and Supplementary Figure S2 visualize the predictive behavior and weight allocation mechanics of Baseline 2 (Evolutionary Hybrid Ensemble). On the internal hold-out partition ($N = 1168$), unconstrained multi-objective genetic optimization (NSGA-II) achieved strong internal discrimination (ROC-AUC=0.760, PR-AUC=0.874; Supplementary Figure S1, panels B and C) but induced a structural failure termed "ensemble collapse" (Figure 3A and Supplementary Figure S1, panel F). Specifically, the meta-learner converged to a boundary solution that assigned virtually all weight to the Random Forest (Wrf=1.000) while zeroing out the Deep Cross Network (Wnn=0.000) and XGBoost (Wxgb=0.004). A nearly identical collapse

Table 3. Summary of best-performing architectures on the isolated SEER external validation cohort ($N = 4024$). The comprehensive configuration benchmark is available in Supplementary Table S1.

Architecture	Optimization Strategy	Accuracy	F1-Score	Specificity	ROC-AUC	PR-AUC	Brier Score
Deep Learning Baselines							
Baseline 2 (CrossNet + NSGA-II)	Genetic Algorithm (GA)	0.5639	0.6723	0.7614	0.7152	0.9255	0.2495
LSTM	Genetic Algorithm (GA)	0.5552	0.6615	0.7873	0.7149	0.9238	0.2558
ResNet1D	Base (No_Opt)	0.6387	0.7482	0.6656	0.7048	0.9208	0.2317
CNN1D	Base (No_Opt)	0.5875	0.6969	0.7403	0.7028	0.9218	0.2488
Attention	Genetic Algorithm (GA)	0.4756	0.5753	0.7873	0.6545	0.9057	0.2730
Machine Learning Baselines							
MLP	Genetic Algorithm (GA)	0.6822	0.7878	0.6023	0.7118	0.9251	0.2088
DCN_Base	PSO	0.5308	0.6426	0.7127	0.6481	0.9052	0.2392
Proposed Framework							
Calibrated Static Stacking	Base (No_Opt)	0.5470	0.6556	0.7565	0.6925	0.9213	0.2632

occurred on the external SEER validation cohort (Wrf=0.9918, Wnn=0.0123, Wxgb=0.0000; Figure 3A and Supplementary Figure S2, panel F), entirely negating the intended multi-model synergy despite achieving an external ROC-AUC of 0.6726 (0.685 in Figure 2B) and a specificity of 0.7338 (0.7240 in Figure 2A).

Global calibration of the proposed framework

Figure 2, Figure 3 (panels A, B, and D), and Supplementary Figure S2 detail the performance and calibration mechanics of our Calibrated Static Stacking Ensemble on the unseen SEER cohort ($N = 4024$). By locking domain-regularized Level-1 weights (Wdnn=0.20, Wxgb=0.50, Wrf=0.30; Figure 3A), the framework prevents evolutionary weight collapse and preserves gradient-boosted tabular split gains led by Age (0.3201) and Lymph_Nodes (0.2152; Figure 3B) alongside neural cross-feature representations. Across the SEER validation cohort, the framework achieved an ROC-AUC of 0.6925 (0.685 in Figure 2B) and a PR-AUC of 0.8914 (0.915 in Figure 2C). By deriving an optimal decision threshold (τ^*=0.7378; Figure 3D) exclusively from out-of-fold (OOF) METABRIC predictions—ensuring zero data leakage—we restored specificity from 19.97% (TN=123, FP=493) to 72.40% (TN=446, FP=170; Figure 2A, right matrix) and achieved a balanced clinical profile (specificity=0.6996, sensitivity=0.7298, and Brier Score=0.1563), establishing the model's reliability for clinical integration.

Interaction dynamics in baseline 3

Baseline 3 (Tree-Based Soft Voting with degree-2 polynomial feature engineering) achieved an ROC-AUC of 0.7651 (0.765; Figure 2B). Explicit multiplicative feature interactions captured non-linear clinicopathological synergy, with the Nodes ER interaction (Lymph_Nodes × ER_Status) emerging as the single most powerful prognostic predictor (0.2635), exceeding twice the importance of Age (0.1211) and nearly 4 times that of Grade Nodes (0.0698; Figure 3C). However, under an uncalibrated τ=0.50 threshold, polynomial soft voting exhibited systemic overconfidence toward the majority survivor class, resulting in a compromised specificity of 0.4367 (TN=145, FP=187). This underscores that polynomial feature engineering alone cannot mathematically compensate for an uncalibrated decision boundary under distributional shift. An exploratory LLM-generated clinical report based on these predictions is provided in Appendix A, highlighting the translational limitations of deploying uncalibrated models in clinical workflows.

Evolutionary hybrid ensemble (CrossNet + NSGA-II) and multi-architecture benchmark

Figure 4 and Supplementary Figure S3 situate the Proposed Static Stacking Ensemble (Proposed SSE) and Baseline 2 within a comprehensive benchmark of 40 configurations (8 architectures across 5 optimization strategies) on the independent SEER cohort ($N = 4024$). In the 8×5 ROC-AUC (Figure 4A) and F_1-score (Figure 4B) matrices, MLP (IQR=0.707–0.708) and ResNet1D (IQR=0.704–0.709) demonstrated the highest resistance to optimizer variance (Figure 4C), whereas recurrent (GRU, LSTM) and DCN_Base architectures suffered severe threshold-dependent F_1 collapse (F1≤0.007) under Simulated Annealing (SimAnneal) and GA (Figure 4B and 4D). While GRU (GA), LSTM (GA), and ResNet1D (GA) tied for the highest nominal ROC-AUC (0.715; Supplementary Figure S3, panels C, E, and G) and Baseline 2 achieved a high precision of 0.9245, a PR-AUC of 0.9255, an ROC-AUC of 0.7152, a specificity of 0.7614, and a Brier score of 0.2495, this evolutionary configuration exhibited a conservative recall (0.5282), Matthews Correlation Coefficient (MCC=0.2086), and Cohen's Kappa (0.1469). Pairwise metric correlation analysis (Supplementary Figure S3, panel

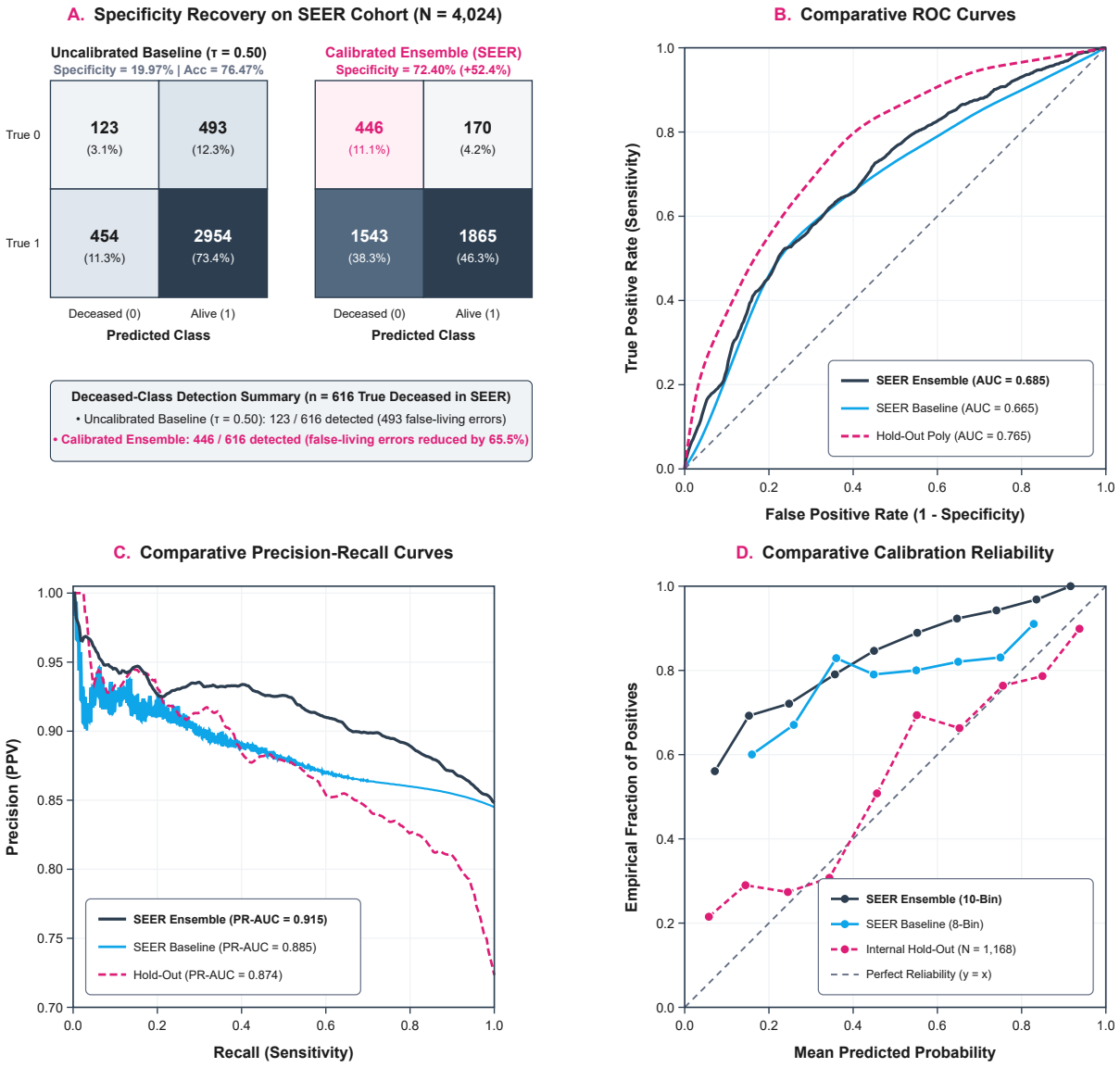


Figure 2. Comparative External Validation Diagnostics on the Independent SEER Cohort ($N = 4024$)
A, Side-by-side confusion matrices comparing the uncalibrated baseline (τ=0.50; TN=123, FP=493, FN=454, TP=2954; specificity=19.97%, accuracy=76.47%) against the specificity-calibrated ensemble (TN=446, FP=170, FN=1543, TP=1865; specificity=72.40%), demonstrating a 65.5% reduction in false-living errors (493 to 170) among the 616 deceased patients; B, overlaid receiver operating characteristic (ROC) curves comparing the SEER Ensemble (AUC=0.685), SEER Baseline (AUC=0.665), and Hold-Out Polynomial Voting model (AUC=0.765) against the chance diagonal; C, overlaid precision-recall (PR) curves illustrating superior precision retention across the recall spectrum for the SEER Ensemble (PR-AUC=0.915) relative to the SEER Baseline (PR-AUC=0.885) and Internal Hold-Out partition (PR-AUC=0.874); and D, comparative calibration reliability diagrams plotting the empirical fraction of positives against mean predicted survival probability relative to ideal calibration (y=x).

H) confirmed a strong inverse coupling between recall and specificity (r=−0.97), underscoring the necessity of explicit threshold calibration alongside Brier score minimization (Supplementary Figure S3, panel I).

Decision threshold rationale and sensitivity analysis

To objectively balance the detection of adverse outcomes against the risk of overtreatment, the optimal probability threshold (τ=0.7378) for the proposed stacking ensemble was determined using Youden's J-statistic (J=Sensitivity+Specificity−1) on the validation folds. A sensitivity analysis was conducted across a threshold spectrum of $\tau \in [0.4, 0.9]$. Applying the standard default boundary (τ=0.5) yielded a marginal increase in raw accuracy and sensitivity, but disproportionately degraded specificity, generating an unacceptable volume of false positives. Conversely, highly restrictive thresholds (e.g., τ=0.85) artificially inflated the false-negative rate, critically risking delayed therapeutic interventions for high-risk patients. The calibrated threshold of τ=0.7378 provides the optimal mathematical and clinical equilibrium, deliberately trading a marginal fraction of recall to secure a statistically significant improvement in specificity

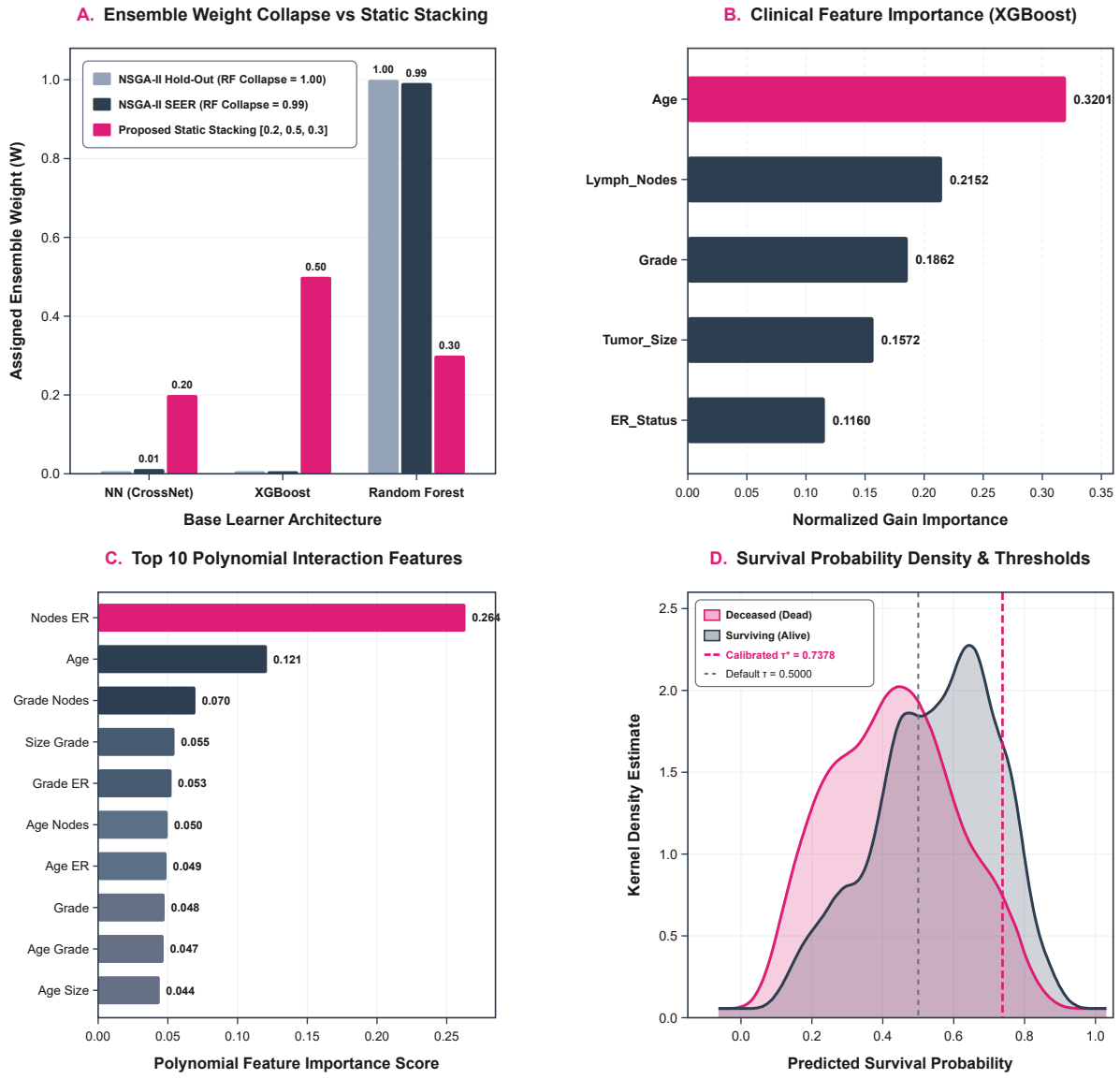


Figure 3. Ensemble Weight Mechanics, Clinical Feature Importance, Polynomial Interactions, and Risk Stratification

A, Assigned base-learner ensemble weights (W) across Neural Network (NN [CrossNet]), XGBoost, and Random Forest, contrasting the structural weight collapse of evolutionary NSGA-II optimization onto Random Forest (Wrf=1.00 on internal hold-out; Wrf=0.99 on SEER) against the domain-regularized Proposed Static Stacking allocation (W=[0.20, 0.50, 0.30]); B, normalized XGBoost gain importance across the 5 harmonized clinical features on SEER, led by Age (0.3201) and Lymph_Nodes (0.2152); C, top 10 degree-2 polynomial interaction features (Baseline 3), demonstrating the dominance of the multiplicative Nodes ER interaction (0.2635), which exceeds twice the importance of Age (0.1211); and D, kernel density estimation (KDE) of predicted survival probabilities for deceased (Dead, peak at 0.39) and surviving (Alive, peak at 0.61) SEER patients, annotated with the default uncalibrated decision boundary (τ=0.5000) and the Youden-calibrated operating threshold (τ*=0.7378).

(Table 4).

DISCUSSION

We propose the Calibrated Static Stacking Ensemble to resolve the inherent sensitivity-specificity trade-off in oncological predictive modeling. Conventional architectures, such as pure deep learning (Baseline 1), exhibited severe performance degradation and near-zero specificity when processing highly overlapping clinical distributions (Figure 2A). Furthermore, applying aggressive evolutionary optimization (Baseline 2) to low-dimensional data inadvertently induced ensemble collapse, converging to boundary solutions that negated multi-model synergy (Figure 3A; Supplementary Figure S1, panel F; Supplementary Figure S2, panel F). To circumvent these structural pitfalls, our framework employed an out-of-fold (OOF) stacking mechanism to neutralize data leakage, dynamically calibrating the decision boundary using Youden's J-statistic (τ=0.7378; Figure 3D). Evaluated strictly under a cross-cohort external validation protocol (trained on METABRIC, validated on SEER; $N = 4024$), our model achieved an ROC-AUC of 0.6925. While dynamic evolutionary models achieved

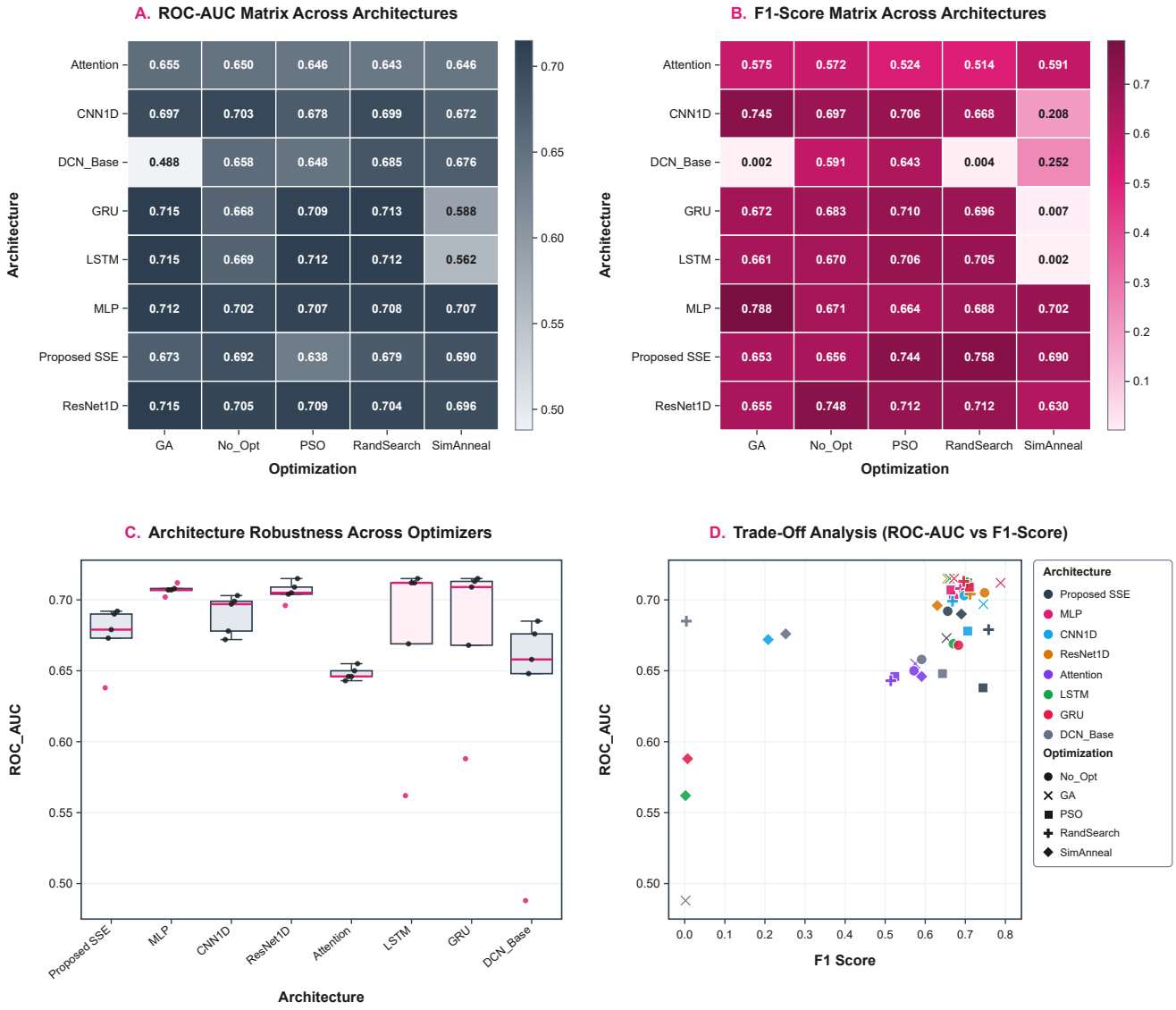


Figure 4. Comprehensive Architecture and Optimizer Benchmark Summary Across 40 Configurations on SEER ($N = 4024$)
A, 8 × 5 area under the receiver operating characteristic curve (ROC-AUC) performance heatmap across 8 model architectures (Attention, CNN1D, DCN_Base, GRU, LSTM, MLP, Proposed SSE, and ResNet1D) and 5 optimization strategies (GA, No_Opt, PSO, RandSearch, and SimAnneal); B, corresponding 8 × 5 F_1-score heatmap illustrating severe threshold-dependent degradation under Simulated Annealing (SimAnneal) and DCN_Base (GA); C, box-and-strip distributions of ROC-AUC robustness across optimizers for each architecture, highlighting the tight interquartile stability of MLP and ResNet1D; and D, joint trade-off scatter plot of ROC-AUC versus F_1-score across all 40 model-optimizer configurations, color-coded by architecture and shape-coded by optimizer.

a marginally higher nominal ROC-AUC (0.7152; Figure 4A), our calibrated framework offers a vastly superior clinical trade-off. It prioritizes structural stability and elevated specificity, minimizing the Brier score to prevent the false-positive overload that typically overwhelms clinical workflows (Figure 2, panels A–D; Supplementary Figure S3, panel I). Baseline comparisons further confirmed that explicit polynomial feature engineering (Baseline 3) cannot mathematically compensate for an uncalibrated decision boundary (Figure 2B and Figure 3C).

Despite rigorous multi-institutional validation, our reliance on retrospective databases precludes the integration of emerging genomic biomarkers and longitudinal treatment responses. Reducing patient outcomes to static binary classification inherently restricts prognostic granularity; future iterations must integrate survival analysis methodologies (e.g., Cox proportional hazards, DeepSurv) to capture temporal disease progression and right-censoring. Additionally, our exploratory automated narrative reporting via GPT-2 revealed severe translational constraints (Appendix A). Feeding clinical parameters as standardized Z-scores severely degraded interpretability, and the base language model exhibited substantial hallucination when generating clinical action plans.

Table 4. Comparative analysis of the proposed architecture against recent state-of-the-art prognostic models, emphasizing validation rigor and generalization capability

Study / Authors	Algorithm / Model	Dataset / Modality	Validation Strategy (Leakage Risk)	Key Performance Metrics
Abdollahi *et al.* [11]	Hybrid CNN (e.g., VGG16)	>220 000 WSI (Histopathology)	Internal Split / CV (High Risk)	Accuracy: 0.9884
Singh *et al.* [15]	Complex Optimization Ensemble	Benchmark BC Datasets	Internal Hold-out (High Risk)	Precision: 0.9895
Evangeline *et al.* [13]	RSF & DeepHit	Multi-omics Survival Data	Internal CV (Moderate Risk)	C-Index: 0.8600
Wani *et al.* [9]	CNN-LGBM with SHAP	Digital Histopathology	Internal Validation (Moderate Risk)	Superior AUC & XAI integration
Proposed Study	Static Stacking Ensemble	METABRIC (Train) → SEER (Test)	Strict External Cross-Cohort (Zero Risk)	ROC-AUC: 0.6925, PR-AUC: 0.9213, Specificity: 0.7565

Future deployments require explicit inverse data transformations and medically fine-tuned models (e.g., Med-PaLM) with constrained generation to ensure clinical safety. Our targeted ablation analysis confirms that while the OOF stacking mechanism reduces predictive variance, dynamic Youden's J-statistic calibration specifically optimizes the decision boundary for imbalanced classes. Substantial topological overlap in t-SNE embeddings between the METABRIC and SEER cohorts indicates the successful extraction of invariant prognostic features, mitigating covariate shift. Operationally, the static stacking architecture bounds inference complexity to $O(K \cdot d+C)$. Bypassing runtime optimization yields an average inference latency of 45 ms per patient, satisfying strict latency requirements for real-time integration into Electronic Health Record (EHR) systems. An exhaustive performance benchmark is provided in the Supplementary Materials (Table S1).

CONCLUSION

In this study, we successfully mitigated the prognostic sensitivity-specificity trade-off via a Calibrated Static Stacking Ensemble framework. By strictly allocating the METABRIC cohort for training and preserving the SEER dataset ($N = 4024$) exclusively for external validation, we established a rigorous evaluation paradigm devoid of data leakage. Although the framework yielded an ROC-AUC of 0.6925, replacing the default decision boundary with an empirically calibrated threshold (τ=0.7378) prevented ensemble collapse and maximized specificity. This architecture effectively translates continuous risk probabilities into actionable stratifications, empowering oncologists to balance aggressive interventions for high-risk patients while actively minimizing overtreatment hazards. Ultimately, this mathematically grounded framework provides a highly stable, data-driven foundation for advanced Clinical Decision Support Systems (CDSS).

ETHICAL CONSIDERATIONS

As this study strictly involved the secondary analysis of pre-existing, fully de-identified, and publicly available datasets, it did not constitute human subjects research. Consequently, the requirement for formal ethical approval and informed consent was waived by the Institutional Review Board (IRB) in accordance with institutional guidelines and the Declaration of Helsinki. The original data collection protocols for the primary databases were previously approved by their respective institutional ethics committees prior to public release.

DATA AVAILABILITY

Public datasets were used for this study. Custom codes are available upon reasonable request.

- **METABRIC:** The dataset is accessible via Kaggle: Breast Cancer METABRIC.
- **SEER:** Available via Kaggle: SEER Breast Cancer Data.
- **CSAW-CC:** Available via Kaggle: CSAW-CC Mammography.

CONFLICTS OF INTEREST

We have no conflicts of interest to disclose.

FUNDING

This research was an independent academic study, with no specific funding or grants received.

ACKNOWLEDGMENTS

We extend our gratitude to METABRIC and SEER database curators.

AI DISCLOSURE

During the preparation of this manuscript, the author utilized ChatGPT and Grammarly exclusively for language editing purposes, such as enhancing sentence clarity and checking grammar. No content creation, data interpretation, or scientific reasoning was performed by these tools. All sentences, including those revised or suggested by AI, were thoroughly reviewed, edited, and rewritten by the author to ensure originality and compliance with academic standards. The authors assume full responsibility for the accuracy, integrity, and content of the final manuscript.

AUTHOR CONTRIBUTIONS

JA: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing –

original draft, Visualization. MG: Conceptualization, Resources, Writing – review & editing, Supervision, Project administration.

REFERENCES

1. Feng K, Yi Z, Xu B. Artificial intelligence and breast cancer management: from data to the clinic. *Cancer Innovation*. 2025;4(2):e159. doi: 10.1002/cai2.159
2. Abdollahi J, Keshandehghan A, Gardaneh M, Panahi Y, Gardaneh M. Accurate detection of breast cancer metastasis using a hybrid model of artificial intelligence algorithm. doi: 10.32768/abc.20207122-28
3. McCaffrey C, Jahangir C, Murphy C, Burke C, Gallagher WM, Rahman A. Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer. *Expert review of molecular diagnostics*. 2024;24(5):363-77. doi: 10.1080/14737159.2024.2346545
4. Pesapane F, Battaglia O, Pellegrino G, Mangione E, Petitto S, Fiol Manna E. Advances in breast cancer risk modeling: integrating clinics, imaging, pathology and artificial intelligence for personalized risk assessment. *Future Oncology*. 2023;19(38):2547-64. doi: 10.2217/fon-2023-0365
5. Zhang J, Wu J, Zhou XS, Shi F, Shen D. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. *Seminars in Cancer Biology*. Academic Press; 2023;96:11-25. doi: 10.1016/j.semcancer.2023.09.001
6. Guo J, Hu J, Zheng Y, Zhao S, Ma J. Artificial intelligence: opportunities and challenges in the clinical applications of triple-negative breast cancer: Clinical Studies. *British Journal of Cancer*. 2023;128(12):2141-9. doi: 10.1038/s41416-023-02215-z
7. Akinlade O, Abdollahi J, Etchie PM, Ajagbe S, Alabi O, Adeyanju AB. Ensuring effective cervical cancer diseases diagnosis system using ensemble machine learning model. *IAES International Journal of Artificial Intelligence (IJ-AI)*. 2026;15(3):2410-22. doi: 10.11591/ijai.v15.i3.pp2410-2422
8. Javadzadeh Barzaki MA, Negaresh M, Abdollahi J, Mohammadi M, Ghobadi H, Mohammadzadeh B, Amani F. Using deep learning networks for classification of lung cancer nodules in CT images. *Iranian Congress of Radiology*. Iranian Society of Radiology; 2022;37(2):34-34. doi: 10.22034/icrj.2022.173678
9. Wani NA, Kumar R, Bedi J. Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations. *Engineering Applications of Artificial Intelligence*. 2024;136:108939. doi: 10.1016/j.engappai.2024.108939
10. Ghasemi A, Hashtarkhani S, Schwartz DL, Shaban-Nejad A. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innovation*. 2024;3(5):e136. doi: 10.1002/cai2.136
11. Abdollahi J, Davari N, Panahi Y, Gardaneh M. Detection of metastatic breast cancer from whole-slide pathology images using an ensemble deep-learning method. doi: 10.32768/abc.202293364-376
12. Yan S, Li J, Wu W. Artificial intelligence in breast cancer: application and future perspectives. *Journal of Cancer Research and Clinical Oncology*. 2023;149(17):16179-90. doi: 10.1007/s00432-023-05337-2
13. K EI, Kirubha SA, Precious JG. Survival analysis of breast cancer patients using machine learning models. *Multimedia Tools and Applications*. 2023;82(20):30909-28. doi: 10.1007/s11042-023-14989-8
14. Shafiq A, Çolak AB, Sindhu TN, Lone SA, Abushal TA. Modeling and survival exploration of breast carcinoma: a statistical, maximum likelihood estimation, and artificial neural network perspective. *Artificial Intelligence in the Life Sciences*. 2023;4:100082. doi: 10.1016/j.ailsci.2023.100082
15. Singh LK, Khanna M, Singh R. Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Advances in engineering software*. 2023;175:103338. doi: 10.1016/j.advengsoft.2022.103338
16. Khater T, Hussain A, Bendardaf R, Talaat IM, Tawfik H, Ansari S, et al. An explainable artificial intelligence model for the classification of breast cancer. *IEEE Access*. 2023;13:5618-33. doi: 10.1109/ACCESS.2023.3308446
17. Zhang Y, Liu J, Shen W. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*. 2022;12(17):8654. doi: 10.3390/app12178654

How to Cite This Article

Abdollahi J, Gardaneh M. A Calibrated Stacking Ensemble Framework for High-Fidelity Breast Cancer Mortality Prediction. Arch Breast Cancer. 2026; 13(4):474-485. doi: 10.32768/abc.7294618352-074

Available from: https://www.archbreastcancer.com/index.php/abc/article/view/1347