Cancer Management and Research

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

A Machine Learning Model for Predicting Breast Cancer Recurrence and Supporting Personalized Treatment Decisions Through Comprehensive Feature Selection and Explainable Ensemble Learning

Tsair-Fwu Lee[1–4], Jun-Ping Shiau[1,5], Chia-Hui Chen[1], Wen-Ping Yun[1], Cheng-Shie Wuu[6], Yu-Jie Huang[7], Shyh-An Yeh[1,8,9], Hui-Chun Chen[7], Pei-Ju Chao[1,7]

[1]Medical Physics and Informatics Laboratory of Electronics Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, 80778, Taiwan, Republic of China; [2]Graduate Institute of Clinical Medicine, Kaohsiung Medical University, Kaohsiung, 807, Taiwan, Republic of China; [3]Department of Medical Imaging and Radiological Sciences, Kaohsiung Medical University, Kaohsiung, 80708, Taiwan, Republic of China; [4]School of Dentistry, College of Dental Medicine, Kaohsiung Medical University, Kaohsiung, 80708, Taiwan, Republic of China; [5]Division of Breast Oncology and Surgery, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, 807, Taiwan, Republic of China; [6]Department of Radiation Oncology, Columbia University, New York, NY, USA; [7]Department of Radiation Oncology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan, Republic of China; [8]Department of Medical Imaging and Radiological Sciences, I-Shou University, Kaohsiung, 82445, Taiwan, Republic of China; [9]Department of Radiation Oncology, E-DA Hospital, Kaohsiung, 82445, Taiwan, Republic of China

Correspondence: Pei-Ju Chao; Hui-Chun Chen, Department of Radiation Oncology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan, Republic of China, Tel +886-7-731-7123 ext 7060, Fax +886-7-732-2813, Email pjchao99@gmail.com; kuas999@gmail.com

Purpose: This study investigates the efficiency of a machine learning model integrating least absolute shrinkage and selection operator (LASSO) feature selection with ensemble learning in predicting recurrence risk and supporting personalized treatment decisions in breast cancer patients.

Materials and Methods: Clinical data from 1,131 breast cancer patients (1,056 nonrecurrent and 75 recurrent) were collected from Kaohsiung Medical University Hospital's electronic health record system. After preprocessing and standardization, LASSO was applied for feature selection. An ensemble learning model was developed based on multiple machine learning algorithms, with SHAP (Shapley additive explanations) used for interpretability.

Results: The ensemble model achieved an AUC of 0.817, outperforming the best single model (AUC 0.711), demonstrating improved predictive accuracy and stability. LASSO identified six key predictors: regional lymph node positivity, ER status, Ki-67, lymphovascular invasion, tumor size, and age at diagnosis. SHAP analysis enhanced transparency by quantifying the contribution of each feature to recurrence risk, improving clinical understanding.

Conclusion: This LASSO-enhanced ensemble model significantly improves the accuracy and interpretability of breast cancer recurrence prediction. By identifying individualized recurrence risks through SHAP analysis, the model supports more precise, data-driven clinical decision-making. These findings demonstrate its potential as a clinical decision support tool for guiding personalized treatment strategies, contributing to more effective breast cancer management.

Plain language summary: Breast cancer is the most common cancer in women worldwide, and despite treatment, some patients experience recurrence, meaning the cancer returns after initial therapy. Identifying which patients are at higher risk of recurrence is crucial for personalized treatment. However, traditional risk prediction models often lack accuracy and do not fully capture the complexity of patient data.

This study developed an ensemble learning model to predict breast cancer recurrence more accurately by integrating LASSO feature selection and multiple machine learning models. Using data from 1,131 breast cancer patients, the model identified six key predictors of recurrence, including lymph node positivity, ER status, Ki-67, lymphovascular invasion, tumor size, and age at diagnosis. The ensemble model achieved higher accuracy (AUC = 0.817) compared to traditional models.

Received: 3 January 2025
Accepted: 19 April 2025
Published: 8 May 2025

To enhance interpretability, SHAP analysis was applied to explain how each factor influences predictions. This transparency helps clinicians understand individualized risk and supports personalized treatment decisions. The model can assist in tailoring treatments—allowing high-risk patients to receive more aggressive care while helping low-risk patients avoid unnecessary treatments.

Future research should focus on validating the model in different populations and incorporating additional data sources like genomics and imaging to further improve precision. This study demonstrates the potential of ensemble learning in advancing personalized breast cancer care.

Keywords: breast cancer recurrence, machine learning, LASSO feature selection, ensemble learning, SHAP value analysis

Introduction

Breast cancer remains the most common malignancy among women, with approximately 2.3 million new cases annually, accounting for over 15% of cancer-related deaths[1]. Despite advancements in diagnosis and treatment, local recurrence remains a challenge, significantly affecting survival and quality of life.[2,3] Breast cancer recurrence can occur many years after initial treatment, especially in patients with larger tumors or lymph node involvement.[4] Accurately identifying high-risk patients is crucial for guiding personalized treatment strategies.

Breast cancer is highly heterogeneous, and the risk of local recurrence varies significantly among patients.[5] Conventional risk models relying on limited clinical and pathological features often lack accuracy and interpretability, especially in complex cases. Managing high-dimensional clinical data while ensuring model robustness remains a key challenge.[6]

Precision medicine has emerged as a strategy to optimize treatment through individualized risk assessment.[7] However, traditional models struggle to integrate diverse clinical and biological factors, leading to suboptimal treatment decisions. High-risk patients may not receive timely interventions, while low-risk patients could be overtreated, increasing side effects and costs.

Machine learning (ML) offers a data-driven approach to modeling complex interactions and improving recurrence risk prediction. However, traditional statistical models face challenges such as overfitting and limited interpretability. The "black box" nature of many ML algorithms further reduces physician trust.[8]

To address these issues, this study develops a breast cancer recurrence risk prediction model that enhances both accuracy and interpretability. We employ LASSO regression for feature selection to identify key clinical variables while mitigating overfitting.[9] A stacking ensemble learning approach integrates multiple ML models to enhance predictive performance and stability.[10]

To improve model interpretability, we incorporate Shapley additive explanations (SHAP) to quantify the contribution of each feature to recurrence risk.[11,12] Unlike traditional feature importance methods, SHAP provides both global (overall risk factors) and local (individual patient influence) interpretability, bridging the gap between ML models and real-world clinical applications.

This model improves recurrence risk prediction while providing clinicians with transparent decision support, facilitating personalized, evidence-based breast cancer treatment.[13] The ultimate goal is to develop a clinical decision support system (CDSS); however, further validation through multi-center trials is necessary to ensure clinical applicability.

Materials and Methods

This study aimed to develop a machine learning-based breast cancer local recurrence risk prediction model by integrating extensive clinical data and advanced data analysis techniques, with the goal of improving risk assessment and supporting personalized treatment decisions. The overall research framework (Figure 1) covers data collection, preprocessing, feature selection, model development, and evaluation to ensure the reliability and clinical applicability of the results.

Data Collection

Clinical data were retrospectively obtained from 1,131 breast cancer patients recorded in the Kaohsiung Medical University Chung Ho Memorial Hospital Cancer Registry. The dataset included 1,056 patients without recurrence and 75 patients with recurrence, diagnosed and treated between 2001 and 2022. This study was conducted in accordance with

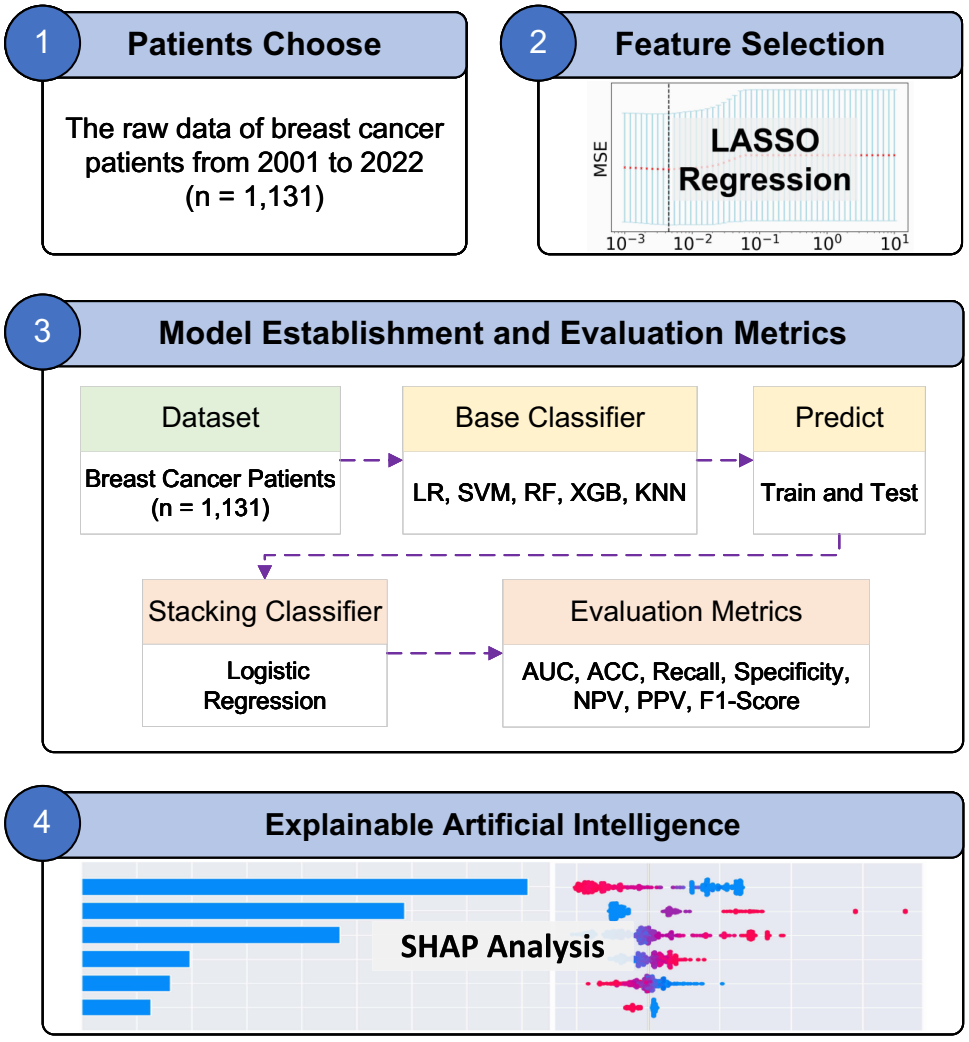

Figure 1 Research flow chart.
Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGB, eXtreme Gradient Boosting; KNN, K-Nearest Neighbors; SHAP, Shapley additive explanation.

the declaration of Helsinki. IRB approval was granted by the Kaohsiung medical university institutional review board (KMUHIRB-E(I)20220126), and the requirement for informed consent was waived due to the retrospective nature of the study, which utilized pre-existing, de-identified clinical data.

Inclusion & Exclusion Criteria

- Inclusion Criteria:
  1. Female patients diagnosed with primary breast cancer.
  2. Underwent mastectomy or lumpectomy as the primary treatment.
  3. Complete medical records with documented recurrence status.
- Exclusion Criteria:
  1. Incomplete clinical records or missing follow-up data.
  2. Patients diagnosed with metastatic breast cancer at baseline.

https://doi.org/10.2147/CMAR.S514693

This study adheres to retrospective cohort study design principles, using only historical, de-identified patient records without additional data collection or direct patient interaction, aligning with ethical considerations.

Variables Collected

A total of 16 clinical variables were collected, covering:

- Demographics: Age at diagnosis, BMI.
- Pathological features: Tumor size, lymph node status, histological grade, molecular subtype (ER, PR, HER2, Ki-67).
- Treatment details: Type of surgery
- Other factors: Smoking behavior, primary tumor site, laterality.

These data form a robust foundation for analyzing breast cancer recurrence risk factors and model development.

Data Preprocessing

To ensure data quality and model robustness, a multi-step data preprocessing strategy was implemented:

1. Standardization: Continuous variables were standardized to a mean of 0 and a standard deviation of 1 to ensure comparability across features. The formula (1) used was:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x represents a data point, μ the mean, and σ the standard deviation.

2. Categorical variable encoding (one-hot encoding): Categorical features were transformed into binary vectors, ensuring that categorical data were processed without implying ordinal relationships.

3. Handling class imbalance (SMOTEENN): Due to the significant imbalance between recurrence (75 cases) and non-recurrence (1,056 cases), resampling was performed using SMOTEENN, a hybrid approach combining:[14,15]

- SMOTE (synthetic minority oversampling technique) to generate synthetic minority class samples.
- Edited nearest neighbors (ENN) to remove redundant majority class samples, refining the decision boundary and reducing noise.

Justification for Choosing SMOTEENN

- Compared to pure SMOTE or random undersampling, SMOTEENN provides superior balance while avoiding over-reliance on synthetic data.[16–18]
- Compared to ADASYN or borderline-SMOTE, which may oversample near decision boundaries, SMOTEENN ensures more robust and representative augmentation.[16–18]
- Final class distribution post-SMOTEENN was adjusted to 0.7: 1 (not fully 1: 1) to enhance model generalization while maintaining sensitivity to minority cases.

SMOTEENN's effectiveness in handling clinical data imbalance enhances predictive performance and model robustness, ensuring that recurrence cases are adequately represented in the training process.

Feature Selection

Feature selection was performed using LASSO (least absolute shrinkage and selection operator) regression, which applies L1 regularization to eliminate irrelevant or redundant variables, preventing overfitting and improving model generalizability.[9]

Feature Set & Selection Process

- Initial Features (16 total):
  - Demographic: Age at diagnosis, BMI
  - Tumor characteristics: Tumor size, histology behavior, grade, laterality, primary site
  - Lymph node and vascular involvement: Regional lymph node positivity, lymphovascular invasion (LVI)
  - Molecular markers: ER (estrogen receptor), PR (progesterone receptor), HER2, Ki-67
  - Treatment & lifestyle: Surgery method, smoking behavior
- LASSO-selected features (Final 6):
  - Regional lymph node positivity
  - ER status
  - Ki-67
  - Lymphovascular invasion (LVI)
  - Tumor size
  - Age at diagnosis

The LASSO penalty parameter (λ) was optimized using 10-fold cross-validation, testing 50 logarithmically spaced λ values (ranging from 10^{-3} to 10^{1}), selecting the one minimizing cross-validation mean squared error (MSE).

LASSO was chosen over stepwise regression and other methods because it:[19]

- Provides stable feature selection, avoiding the sensitivity to random errors observed in stepwise methods.
- Handles multicollinearity effectively by shrinking redundant predictors to zero, which simplifies the model and improves interpretability.

Machine Learning Model Development

Baseline Models

Five commonly used ML models were implemented:

- Logistic regression (LR)
- Support vector machine (SVM)
- Random forest (RF)
- Extreme gradient boosting (XGB)
- K-Nearest neighbors (KNN)

The dataset was randomly split into 70% training and 30% testing, with 10-fold cross-validation used for hyperparameter tuning.

Integrated Learning Model

To further improve prediction accuracy and model stability, this study used ensemble learning techniques. Ensemble methods include bagging, boosting, and stacking.[20,21] These methods reduce the limitations of a single model by integrating the prediction results of multiple models, thereby generating a more powerful prediction model.

The ensemble learning architecture is shown in Figure 2. The training set data is input into the base classifiers for training, and the predicted probabilities from each classifier are obtained. The dataset was initially divided into training and testing sets in a 7:3 ratio, and the meta-classifier was trained using the predicted probabilities from the base classifiers as input. The final prediction result is generated after training the meta-classifier.

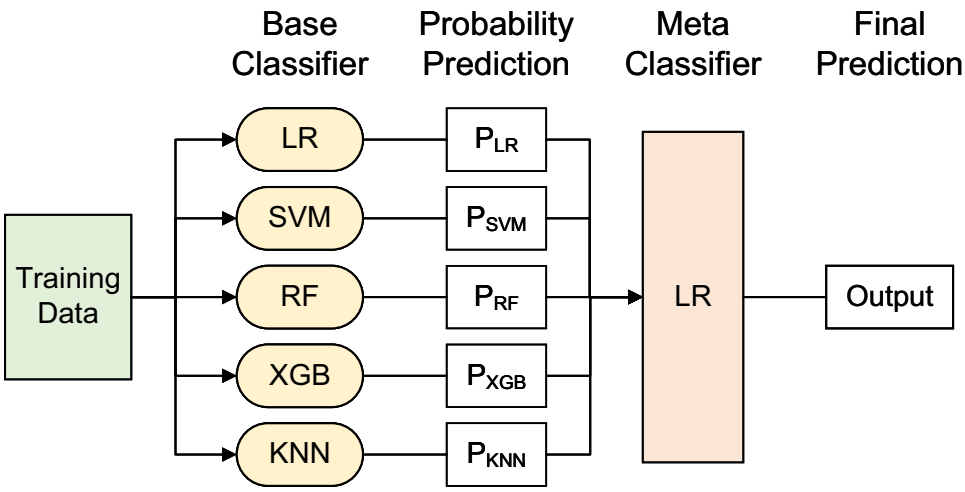

Figure 2 Flowchart of ensemble learning.
Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGB, eXtreme Gradient Boosting; KNN, K-Nearest Neighbors.

To improve model performance and stability, we employed an ensemble learning approach utilizing stacking, which integrates multiple base models into a meta-classifier:

- First layer (base models): Predictions were generated using LR, SVM, RF, XGB, and KNN.
- Second layer (meta-classifier): The meta-classifier was trained on the base model outputs to produce final predictions.

Model Evaluation

To comprehensively assess model performance, we employed the following metrics:

- Area under the curve (AUC) – Measures overall classification performance.
- Accuracy (ACC) – Evaluates overall prediction correctness.
- Recall (sensitivity) – Ensures effective detection of recurrent cases.
- 10-fold cross-validation – Used to reduce bias and confirm generalizability.

SHAP Value Analysis

To enhance clinical interpretability, we applied Shapley additive explanations (SHAP) analysis, which provides:

1. Global Interpretability: Identifies which features contribute most to recurrence prediction.
2. Local Interpretability: Assesses how features impact individual patient predictions.

SHAP outperforms Feature Importance and LIME by offering:

- Consistent, theoretically justified explanations (based on cooperative game theory).
- Both global and local interpretability, enabling personalized clinical decision-making.
- More robust feature explanations, reducing instability compared to alternative methods.

Since LASSO already filtered redundant features, SHAP was applied only to the final selected features to provide clear and clinically meaningful insights.

Results

This study successfully combined advanced machine learning techniques with comprehensive clinical data to develop effective predictive models for breast cancer local recurrence. The clinical data of 1,056 patients without recurrence and 75 patients with recurrence (Table 1) were analyzed, including variables such as age at diagnosis, BMI, and tumor size. These variables not only revealed key risk factors for breast cancer recurrence but also emphasized the significant correlation between tumor size and recurrence risk, further highlighting the importance of careful monitoring.

Table 1 Clinical Characteristics of Patients

Patient Characteristics	No Recurrence (%)	Recurrence (%)	p
	n = 1056	n = 75	
Diagnosis Age			0.065
Mean	53	50	
BMI			0.954
Mean	23.90	23.93	
Tumor Size			0.001
Mean	17.77	21.55	
Regional Lymph Nodes Positive			<0.001
Mean	0.27	1.11	
Smoking Behavior			0.259
No	1015 (89.7)	74 (6.5)	
Yes	41 (3.6)	1 (0.1)	
Primary Site			0.585
C501	116 (10.3)	9 (0.8)	
C502	150 (13.3)	9 (0.8)	
C503	113 (10.0)	5 (0.4)	
C504	461 (40.8)	34 (3.0)	
C505	103 (9.1)	10 (0.9)	
C508	87 (7.7)	8 (0.7)	
C509	26 (2.3)	0 (0.0)	
Histology Behavior			0.093
DCIS	21 (1.9)	1 (0.1)	
Ductal carcinoma	963 (85.1)	69 (6.1)	
Lobular carcinoma	48 (4.2)	2 (0.2)	
Combine ductal and lobular	6 (0.5)	0 (0.0)	
Paget disease	8 (0.7)	3 (0.3)	
Metaplastic carcinoma	10 (0.9)	0 (0.0)	

(*Continued*)

Table 1 (Continued).

Patient Characteristics	No Recurrence (%)	Recurrence (%)	p
	n = 1056	n = 75	
Laterality			0.860
Right	518 (45.8)	36 (3.2)	
Left	538 (47.6)	39 (3.4)	
Surgery Method			0.767
Lumpectomy	601 (53.1)	44 (3.9)	
Mastectomy	455 (40.2)	31 (2.7)	
LVI			0.444
Negative	775 (68.5)	52 (4.6)	
Positive	281 (24.8)	23 (2.0)	
Grade			0.015
1	96 (8.5)	2 (0.2)	
2	581 (51.4)	35 (3.1)	
3	379 (33.5)	38 (3.4)	
ER Status			<0.001
Mean	69.48	46.21	
PR Status			0.012
Mean	45.64	33.60	
ki-67			<0.001
Mean	22.62	34.35	
HER2 Status			0.200
Negative	131 (11.6)	8 (0.7)	
1+	342 (30.2)	19 (1.7)	
2+	354 (31.3)	24 (2.1)	
3+	229 (20.2)	24 (2.1)	
Molecular subtype			<0.001
Luminal A	64 (5.7)	2 (0.2)	
Luminal B1	58 (5.1)	4 (0.4)	
Luminal B2	738 (65.3)	38 (3.4)	
HER2-enriched	187 (16.5)	29 (2.6)	
Triple-negative	9 (0.8)	2 (0.2)	

Abbreviations: BMI, Body Mass Index; DCIS, Ductal Carcinoma In Situ LVI, Lymphovascular Invasion; ER, Estrogen Receptor; PR, Progesterone Receptor; HER2, Human Epidermal Growth Factor Receptor 2.

Key Predictors

A total of 16 clinical characteristics were initially considered for model development. Using LASSO regression for feature selection, we identified six key features that were most predictive of breast cancer local recurrence (Figure 3a). These features, ranked by their importance to recurrence prediction, were: Regional lymph node positivity, ER status, LVI, Ki-67 index, Age at diagnosis, Tumor size.

These features are critical in predicting local recurrence due to their strong correlation with tumor behavior. For example, regional lymph node positivity indicates potential cancer spread beyond the breast, ER status affects hormone therapy responsiveness, and Ki-67 reflects tumor proliferation rates, all of which are known indicators of aggressive cancer.

Model Performance Evaluation

Figure 3b demonstrates the performance comparison between individual models and the ensemble model, with different feature combinations evaluated using the AUC (area under the curve) metric. The ensemble model consistently outperformed the individual models, particularly when using the LASSO-selected features. The ensemble model achieved an AUC of 0.817, which was significantly higher than the best-performing individual model with an AUC of 0.711. This highlights the superior accuracy and stability of the ensemble model in predicting local recurrence.

Table 2 provides a detailed performance comparison of the five different models (logistic regression, support vector machine, random forest, extreme gradient boosting, and k-nearest neighbors) when using all features and LASSO-

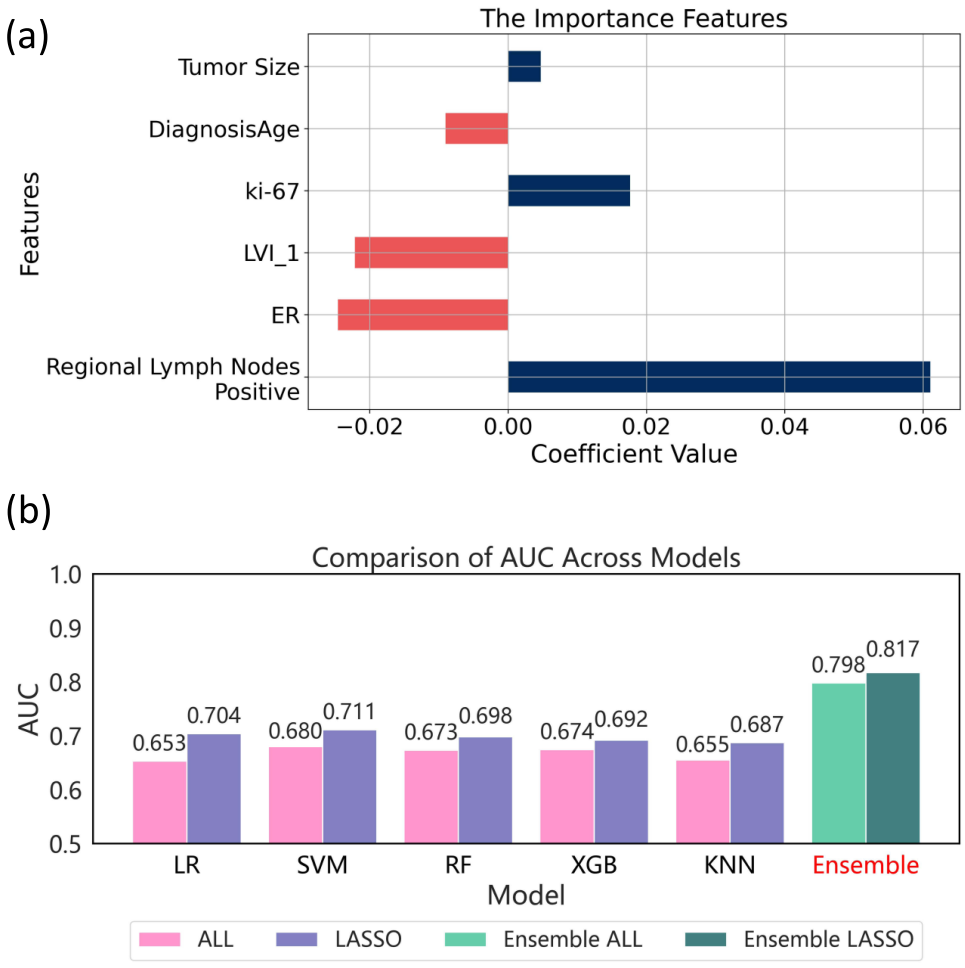

Figure 3 (**a**) Ranking diagram of the importance of feature factors of LASSO. The ranking of important factors selected by LASSO is as follows: Regional Lymph Nodes Positive, ER, LVI, ki-67, Diagnosis Age, Tumor Size; (**b**) Comparison of the AUC of a single model and the integration of all the features and the LASSO selected features. **Abbreviations**: BMI, Body Mass Index; LVI, Lymphovascular Invasion; ER, Estrogen Receptor; PR, Progesterone Receptor; HER2, Human Epidermal Growth Factor Receptor 2; LASSO, Least Absolute Shrinkage and Selection Operator; LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGB, eXtreme Gradient Boosting; KNN, K-Nearest Neighbors.

https://doi.org/10.2147/CMAR.S514693

Table 2 Integrated Model Evaluation Indicators of the Five Models Created by the Combination of All the Features and the LASSO Features

Feature Combination	Model	AUC	ACC	Recall	Specificity	NPV	PPV	F1-Score
All (16)	LR	0.653	0.619	0.453	0.785	0.593	0.669	0.536
	SVM	0.680	0.630	0.443	0.818	0.601	0.693	0.533
	RF	0.673	0.618	0.504	0.732	0.599	0.650	0.565
	XGB	0.674	0.614	0.548	0.680	0.604	0.629	0.583
	KNN	0.655	0.625	0.534	0.715	0.611	0.650	0.582
	Ensemble	0.798	0.718	0.688	0.748	0.705	0.732	0.709
LASSO selected	LR	0.704	0.627	0.499	0.755	0.610	0.662	0.561
	SVM	0.711	0.632	0.448	0.816	0.600	0.703	0.543
	RF	0.698	0.618	0.520	0.717	0.604	0.640	0.571
	XGB	0.692	0.627	0.571	0.684	0.617	0.641	0.603
	KNN	0.687	0.635	0.599	0.671	0.633	0.641	0.616
	Ensemble	0.817	0.732	0.716	0.748	0.725	0.739	0.728

Notes: All **16** Features: These are the full set of clinical features used in the initial model development before applying LASSO feature selection: Diagnosis Age, BMI (Body Mass Index), Tumor Size, Regional Lymph Nodes Positive, Smoking Behavior, Primary Site, Histology Behavior, Laterality, Surgery Method, LVI (lymphovascular invasion), Grade, Molecular Subtype, ER (Estrogen Receptor), PR (progesterone receptor), ki-67, HER2 (Human Epidermal Growth Factor Receptor 2). After applying LASSO feature selection, the most predictive clinical variables selected are: Regional Lymph Nodes Positive, ER (Estrogen Receptor), ki-67, LVI (Lymphovascular Invasion), Tumor Size, Diagnosis Age.

Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGB, eXtreme Gradient Boosting; KNN, K-Nearest Neighbors; AUC, Area Under the Receiver Operating Characteristic Curve; ACC, Accuracy; NPV, Negative Predictive Value; PPV, Positive Predictive Value (Precision).

selected features. After applying the LASSO feature selection, the ensemble model's accuracy improved to 73%, with balanced performance in terms of AUC, accuracy (ACC), and recall. These results underscore the ensemble model's potential for clinical decision support due to its reliability and balanced performance metrics.

SHAP Value Analysis

To enhance the interpretability of the model, SHAP value analysis was conducted to evaluate the contribution of each feature to the model's predictions. SHAP values provide a framework for understanding how individual features impact each prediction, thereby offering greater transparency.

Figure 4a presents the average SHAP values for each LASSO-selected feature, illustrating the importance of regional lymph node positivity, ER status, and Ki-67 as the most influential factors in recurrence prediction. Figure 4b provides a bee colony plot showing the SHAP values for individual samples across various feature ranges, highlighting how these features interact and their specific impact on prediction outcomes. For instance, patients with high SHAP values for regional lymph node positivity were more likely to experience recurrence, while those with lower SHAP values had reduced risk.

By using SHAP analysis, actionable insights can be drawn from the model's predictions. For example, in one patient case with high Ki-67 and positive regional lymph nodes, the SHAP value analysis revealed that these factors significantly increased the predicted recurrence risk, informing the clinical decision to recommend more aggressive treatment. This case highlights how SHAP analysis makes the model more interpretable and useful for personalized treatment planning.

Feature Combination and Model Optimization

The breast cancer local recurrence prediction model developed in this study demonstrated high predictive accuracy and stability, particularly when combining LASSO feature selection with ensemble learning techniques. The ensemble

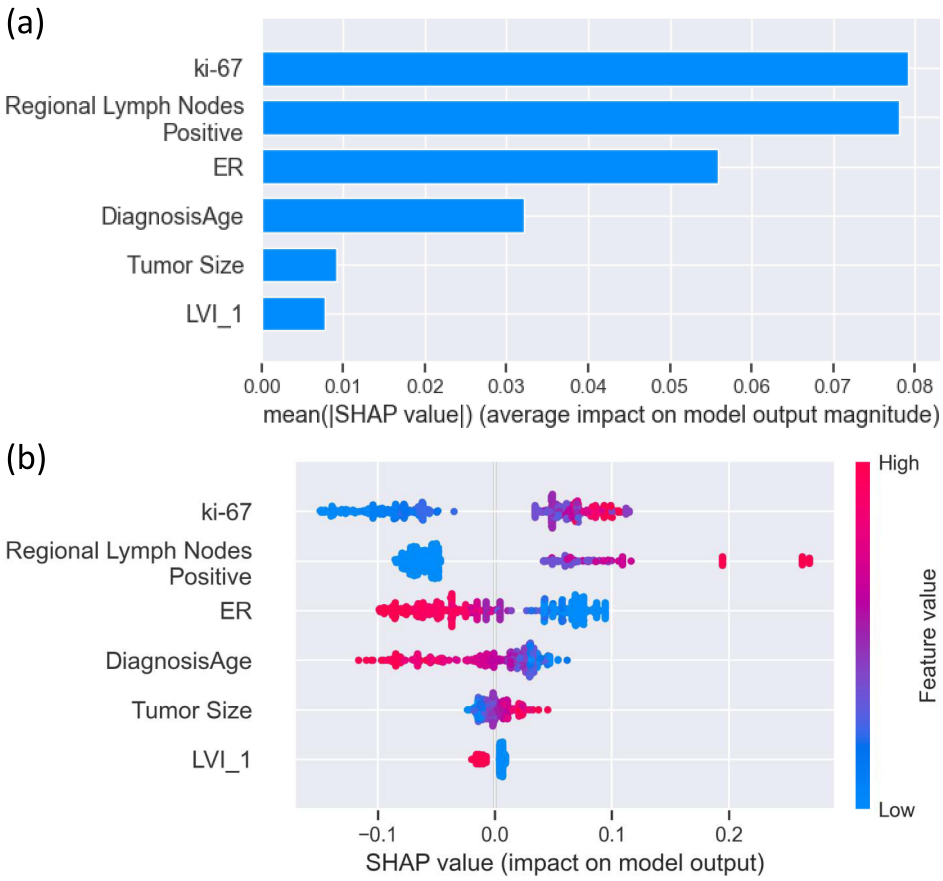

Figure 4 (**a**) Bar plot and (**b**) Beeswarm plot of feature importance in the ensemble machine learning model.
Abbreviations: ER, Estrogen Receptor; LVI, Lymphovascular Invasion.

approach, combined with key clinical features, significantly improved the model's performance. Additionally, SHAP value analysis enhanced the model's interpretability, making it more transparent and applicable for clinical use. These findings not only offer a strong scientific foundation for clinical decision-making but also reinforce the model's potential for real-world medical applications by providing clinicians with an effective tool to identify high-risk patients and develop personalized treatment strategies.

Discussion

Key Findings and Contributions

This study developed a local risk prediction model for breast cancer recurrence by combining LASSO feature selection and ensemble learning technology. The model demonstrated significant predictive performance, especially in improving accuracy and model interpretability. The key results included the following: (1) Six key clinical variables were screened by LASSO and ranked in order of importance: regional lymph node positivity, ER status, lymphovascular invasion (LVI), ki-67, and diagnostic age. (2) Ensemble learning technology significantly improved the stability and accuracy of the model, with the AUC value reaching 0.817. (3) SHAP analysis enhanced the interpretability of the model, enabling clinicians to better understand the significance of the prediction result logic to develop a more personalized treatment plan. These technological innovations demonstrate the potential application value of this model in the personalized treatment of breast cancer.

The six factors screened by LASSO cover many aspects of breast cancer recurrence risk, including the biological behavior of the tumor (ER, ki-67), indicators of cancer spread (positive regional lymph nodes, LVI), size and stage of the tumor (tumor size), and individual characteristics (age at diagnosis) of the patients. The analysis of the clinical significance of these six important factors is explained below.

Positive regional lymph nodes are among the most important prognostic indicators of breast cancer.[22] When cancer cells metastasize to lymph nodes, the risk of recurrence and distant metastasis increases significantly. Patients with positive lymph nodes usually require more aggressive treatment, such as adjuvant chemotherapy or radiotherapy, to reduce the risk of recurrence. The importance of this variable is that it directly reflects the extent of cancer spread and predicts a poor prognosis.[23]

The estrogen receptor (ER) status is the most common hormone receptor status in breast cancer patients, and it is usually associated with a better treatment response and a lower risk of recurrence.[24] ER-positive patients respond better to endocrine therapy (such as tamoxifen or aromatase inhibitors). The ER status plays a key role in treatment decisions and is closely related to the biological behavior of the disease.[25]

Lymphovascular invasion (LVI) refers to the invasion of cancer cells into lymphatic vessels or blood vessels and is an early sign of tumor metastasis. LVI-positive patients have a greater risk of recurrence because cancer cells may metastasize to other locations through the lymphatic or vascular system. Therefore, LVI is an important predictor for determining the need for adjuvant therapy.[26]

ki-67 is a cell proliferation marker that reflects the proliferation rate of tumor cells. High ki-67 levels usually predict faster tumor growth and a greater risk of recurrence and are usually associated with poor prognosis. Clinically, ki-67 is used to predict whether patients need more aggressive treatment strategies. High ki-67 usually indicates that a tumor has more invasive characteristics.[27]

Age at diagnosis is an important variable for the prognosis of patients with breast cancer. Young patients (usually less than 40 years old) usually have more invasive breast cancer and a higher risk of recurrence. In contrast, tumors in elderly patients usually grow slowly, but their treatment tolerance is lower; therefore, individualized treatment is needed. Age at diagnosis helps to determine the intensity of treatment and the evaluation of prognosis.[28]

Tumor size is an important factor for the prognosis of patients with breast cancer. Larger tumors are generally associated with a greater risk of recurrence and poorer prognosis.[22] Patients with tumors larger than 2 cm are considered at greater risk and may require more aggressive treatment, such as adjuvant chemotherapy or radiotherapy after surgery. Tumor size reflects the invasiveness and progression stage of cancer at the time of diagnosis.[29]

Analysis of the Key Influencing Factors

In this study, the most important influencing factors were regional lymph node positivity and ER status. These features have been confirmed as important predictors of breast cancer recurrence in most studies[30,31] and had the highest weight in the LASSO feature selection process in the present study; in particular, positive regional lymph nodes contributed the most to model prediction. In contrast, other variables, such as ki-67, tumor size, diagnosis age and lymphovascular invasion (LVI), still play important roles in prediction, but their relative weights are low, indicating that the impact on the final forecast result is relatively small.

Highlights

This study used LASSO feature selection technology to screen six clinical features with significant predictive power for the risk of local recurrence of breast cancer, which significantly improved the accuracy of the model and reduced the risk of overfitting. The combination of ensemble learning technology further improved the stability and performance of the model and finally reached an AUC value of 0.817, which was significantly better than those of the other single models. Moreover, SHAP value analysis was introduced in the present study, which enhances the transparency and interpretability of the model, enables clinicians to more clearly understand the logic of the prediction results, and is helpful for clinical application. This predictive model can help clinicians accurately identify high-risk patients to develop individualized treatment strategies and promote the development of personalized treatments for patients with potential clinical practicability.

Comparison With Existing Literature

In comparison with existing studies, the innovation of the present study lies in the significant improvement in the accuracy and stability of breast cancer recurrence risk prediction through the combination of LASSO feature selection and ensemble learning techniques. While many studies have already established the correlation between individual

factors, such as ER status, positive regional lymph nodes, LVI, Ki-67, tumor size, and age at diagnosis, and breast cancer recurrence,[32–34] this study is unique in integrating these factors into a machine learning framework. Furthermore, this study goes beyond simply identifying these factors by quantifying their contribution to model predictions using SHAP analysis. This approach not only highlights the predictive power of individual variables like ER status and lymph node positivity but also provides transparency and explainability that were lacking in previous models.

Previous studies have also explored the integration of LASSO-based feature selection with ensemble learning for breast cancer recurrence prediction. For example, Lee et al (2023) developed a radiomics-based machine learning model to predict locoregional recurrence (LRR) using LASSO for feature selection and an ensemble stacking approach, achieving an AUC of 0.78, which was higher than individual models (0.61–0.70).[35] This aligns with our findings, as our LASSO-selected features combined with ensemble learning yielded an AUC of 0.817, further demonstrating the advantage of integrating LASSO with ensemble techniques for improving predictive performance and stability.

Unlike traditional statistical models, which often struggle with high-dimensional clinical data, this study demonstrates the capability of ensemble learning to process such data efficiently while yielding actionable clinical predictions. Previous studies have demonstrated the effectiveness of ensemble learning in medical AI.[10,36] By integrating multiple classifiers, ensemble models improve both predictive accuracy and robustness, making them especially valuable for analyzing complex clinical datasets. Moreover, the use of SHAP analysis allows clinicians to understand how specific factors contribute to the risk of recurrence on an individual patient basis, offering insights that were previously unavailable. This combination of predictive accuracy and interpretability marks a significant advancement over prior studies, providing a more powerful and clinically relevant prediction framework.

Analysis of Key Figures or Tables

Diagram analysis in this study revealed that the ensemble learning model had a significant advantage in processing and assessing the risk of breast cancer recurrence. In particular, the AUC curve shown in Figure 3b reveals that the AUC value of the ensemble model reaches 0.817, which is significantly higher than the performance of the other single models (the highest AUC value is 0.711). In addition, the effects of different feature combinations on the model are shown in detail in Table 2, which validates the ability of the feature set screened by LASSO to improve the prediction accuracy.

Linking With Practical Applications

The design of this prediction model fully addresses the practical needs of clinical applications, particularly in terms of interpretability. The model allows clinicians to understand the logic behind the predictions, which is critical for clinical decision-making. The introduction of SHAP values enhances transparency, helping doctors quickly identify high-risk patients and facilitating the implementation of personalized treatment strategies. Moreover, the model's ability to handle large volumes of heterogeneous clinical data makes it suitable for large-scale breast cancer risk prediction.

By using SHAP analysis, this study significantly improves clinical decision-making by making the model's predictions more interpretable SHAP values break down the overall prediction, revealing the individual contribution of features such as ER status, lymph node positivity, and Ki-67 to a specific patient's recurrence risk. For instance, in a patient with high regional lymph node positivity and elevated Ki-67 levels, SHAP values highlight the dominant role these factors play in the predicted high recurrence risk.

This interpretability is crucial for personalized treatment, as it enables doctors to tailor interventions based on a patient's unique risk profile. Understanding the exact contribution of each risk factor allows clinicians to make informed decisions on whether to escalate treatment, increase monitoring, or adjust therapy. The ability to translate complex predictions into clear, actionable insights supports precision medicine, leading to more targeted and effective treatments.

Research Limitations and Challenges

Although this study achieved good prediction results, it is still limited. First, the data in this study came from a single hospital and may have geographic and population specificity, which affects the generalizability of the model. Second, the sample size of the present study is relatively limited, especially in terms of data from relapsed patients, which may affect the application of the model in larger datasets. Future studies should consider increasing the sample size and carrying out

multicenter cooperation to improve the extensive applicability of the model. In addition, although the SHAP value enhances the interpretability of the model, it is still a postmodel explanation method. In the future, we can explore the construction of a model framework with interpretability itself. Another important factor to consider is that advancements in breast cancer surgical techniques and treatment strategies from 2001 to 2022 may have influenced recurrence risk. Changes in systemic therapies, radiotherapy techniques, and surgical approaches over time could introduce variability in recurrence patterns, which this study could not fully account for. Future research should consider stratifying patients by treatment era to assess the impact of evolving clinical practices on recurrence prediction. Lastly, before clinical implementation, real-world validation through large-scale prospective studies and multi-center trials is essential. Ensuring the model's robustness across diverse healthcare settings will confirm its generalizability and clinical utility. Without external validation, the model risks being overfitted to a specific dataset, limiting its broader applicability. Future research should prioritize external validation to establish its reliability in real-world clinical practice.

Future Research Directions

Future research directions include the following: (1) expand data sources and carry out multicenter data collection to improve the generalizability of the model; (2) introduce multiomics data such as genetic data and image data to further improve the accuracy of prediction; (3) explore more machine learning technologies based on interpretability so that model construction has transparency in the process rather than relying on later explanation methods; and (4) explore more advanced deep learning technologies and reinforcement learning technologies to further improve the performance of the model in dealing with heterogeneous data. (5) Compare resampling techniques, such as ADASYN and Borderline-SMOTE, to improve class balance and model generalizability.

Conclusion

This study successfully developed a machine learning model integrating LASSO feature selection with ensemble learning, significantly improving the prediction accuracy of local recurrence risk in breast cancer patients. By leveraging the strengths of multiple models, the ensemble learning approach demonstrated superior stability and predictive performance compared to individual models, achieving an AUC of 0.817. These findings underscore the effectiveness of multimodel integration in handling complex clinical data and enhancing predictive reliability.

Beyond predictive accuracy, this study also prioritized interpretability through SHAP analysis, enabling clinicians to understand the contribution of each feature to recurrence risk predictions. This interpretability facilitates the identification of high-risk patients and supports the development of personalized treatment strategies, reinforcing the model's potential as a decision-support tool in clinical practice. The ability to provide both robust predictive performance and explainability strengthens its applicability in real-world oncology settings.

Future research should aim to enhance the model's generalizability by incorporating additional data sources, such as genomic and imaging data, and validating its performance across multicenter datasets. Additionally, further exploration of advanced ensemble learning techniques or deep learning architectures may further optimize the accuracy of breast cancer recurrence risk prediction, ultimately supporting more precise and personalized treatment strategies for breast cancer management.

Data Sharing Statement

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request. Because of legal restrictions and ethics, the data in this manuscript are available upon formal request from the corresponding author.

Ethical Approval and Consent to Participate

This study received approval from the institutional review board (IRB) of Kaohsiung medical university (approval number: KMUHIRB-E(I)20220126) and was conducted in accordance with the declaration of Helsinki. The requirement for informed consent was waived due to the retrospective nature of the research, which utilized pre-existing de-identified clinical data.

Acknowledgments

This study was partially supported by a grant from the National Science and Technology Council (NSTC), Executive Yuan, Taiwan, Republic of China (113-2221-E-992-011-MY2). Part of these results was presented in abstract form at the 2024 GBCC Conference.

Funding

Grant from the National Science and Technology Council (NSTC) of the Executive Yuan of the Republic of China, (113-2221-E-992-011-MY2).

Disclosure

All the authors declare that no competing interests exist.

References

1. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast*. 2022;66:15–23. doi:10.1016/j.breast.2022.08.010
2. Group EBCTC. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet*. 2000;355(9217):1757–1770. doi:10.1016/S0140-6736(00)02263-7
3. Bull AA, Meyerowitz BE, Hart S, Mosconi P, Apolone G, Liberati A. Quality of life in women with recurrent breast cancer. *Breast Cancer Res Treat*. 1999;54(1):47–57. doi:10.1023/A:1006172024218
4. Pedersen RN, Esen BÖ, Mellemkjær L, et al. The incidence of breast cancer recurrence 10-32 years after primary diagnosis. *JNCI J National Cancer Inst*. 2022;114(3):391–399. doi:10.1093/jnci/djab202
5. Guo L, Kong D, Liu J, et al. Breast cancer heterogeneity and its implication in personalized precision therapy. *Exp Hematol Oncol*. 2023;12(1):3.
6. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat*. 2012;133(1):1–10. doi:10.1007/s10549-011-1853-z
7. Wang RC, Wang Z. Precision medicine: disease subtyping and tailored treatment. *Cancers*. 2023;15(15):3837. doi:10.3390/cancers15153837
8. Von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. *Philosophy Technol*. 2021;34(4):1607–1622. doi:10.1007/s13347-021-00477-0
9. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Statist Soc B*. 1996;58(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
10. Kumar M, Singhal S, Shekhar S, Sharma B, Srivastava G. Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*. 2022;14(21):13998. doi:10.3390/su142113998
11. Sundararajan M, Najmi A. The many Shapley values for model explanation. *PMLR*. 2020;2020:9269–9278.
12. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed*. 2022;214:106584. doi:10.1016/j.cmpb.2021.106584
13. Chen W, Lu Y, Qiu L, Kumar S. Designing personalized treatment plans for breast cancer. *Inf Syst Res*. 2021;32(3):932–949. doi:10.1287/isre.2021.1002
14. Lee PH. Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *Int J Environ Res Public Health*. 2014;11(9):9776–9789. doi:10.3390/ijerph110909776
15. Muntasir Nishat M, Faisal F, Jahan Ratul I, et al. A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. *Sci Programm*. 2022;2022(1):3649406. doi:10.1155/2022/3649406
16. Gurcan F, Soylu A. Learning from Imbalanced Data: integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers*. 2024;16(19):3417. doi:10.3390/cancers16193417
17. Puri A, Gupta MK. Comparative analysis of resampling techniques under noisy imbalanced datasets. *IEEE*. 2019;2019:1–5.
18. Elsobky AM, Keshk AE, Malhat MG. A comparative study for different resampling techniques for imbalanced datasets. *IJCI Int J Computers Inform*. 2023;10(3):147–156. doi:10.21608/ijci.2023.236287.1136
19. Muthukrishnan R, Rohini R. LASSO: a feature selection technique in predictive modeling for machine learning. *IEEE*. 2016;2016:18–20.
20. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. *MDPI*. 2023;2023:1808.
21. Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*. 2022;10:99129–99149. doi:10.1109/ACCESS.2022.3207287
22. Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*. 1989;63(1):181–187. doi:10.1002/1097-0142(19890101)63:1<181::AID-CNCR2820630129>3.0.CO;2-H
23. Harris EE, Hwang WT, Seyednejad F, Solin LJ. Prognosis after regional lymph node recurrence in patients with stage I–II breast carcinoma treated with breast conservation therapy. *Cancer: Interdiscip Int J Am Cancer Soc*. 2003;98(10):2144–2151. doi:10.1002/cncr.11767
24. Sleightholm R, Neilsen BK, Elkhatib S, et al. Percentage of hormone receptor positivity in breast cancer provides prognostic value: a single-institute study. *J Clin Med Res*. 2021;13(1):9. doi:10.14740/jocmr4398
25. Reinert T, Cascelli F, Resende C, Gonçalves AC, Godo VSP, Barrios CH. Clinical implication of low estrogen receptor (ER-low) expression in breast cancer. *Front Endocrinol*. 2022;13:1015388. doi:10.3389/fendo.2022.1015388
26. Kuhn E, Gambini D, Despini L, Asnaghi D, Runza L, Ferrero S. Updates on lymphovascular invasion in breast cancer. *Biomedicines*. 2023;11(3):968. doi:10.3390/biomedicines11030968

27. Lee J, Lee Y-J, Bae SJ, et al. Ki-67, 21-gene recurrence score, endocrine resistance, and survival in patients with breast cancer. *JAMA Netw Open*. 2023;6(8):e2330961–e2330961. doi:10.1001/jamanetworkopen.2023.30961
28. Cathcart-Rake EJ, Ruddy KJ, Bleyer A, Johnson RH. Breast cancer in adolescent and young adult women under the age of 40 years. *JCO Oncol Pract*. 2021;17(6):305–313. doi:10.1200/OP.20.00793
29. Montero A, Ciervide R, Garcia-Aranda M, Rubio C. Postmastectomy radiation therapy in early breast cancer: utility or futility? *Crit rev oncol/hematol*. 2020;147:102887. doi:10.1016/j.critrevonc.2020.102887
30. Guo J, Fung BC, Iqbal F, et al. Revealing determinant factors for early breast cancer recurrence by decision tree. *Inform Syst Front*. 2017;19(6):1233–1241. doi:10.1007/s10796-017-9764-0
31. Carreño G, Del Casar JM, Corte MD, et al. Local recurrence after mastectomy for breast cancer: analysis of clinicopathological, biological and prognostic characteristics. *Breast Cancer Res Treat*. 2007;102(1):61–73. doi:10.1007/s10549-006-9310-0
32. McCready DR, Hanna W, Kahn H, et al. Factors associated with local breast cancer recurrence after lumpectomy alone. *Ann Surg Oncol*. 1996;3(4):358–366. doi:10.1007/BF02305665
33. Cianfrocca M, Goldstein LJ. Prognostic and predictive factors in early-stage breast cancer. *Oncologist*. 2004;9(6):606–616. doi:10.1634/theoncologist.9-6-606
34. Group EBCTC. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet*. 2005;366(9503):2087–2106. doi:10.1016/S0140-6736(05)67887-7
35. Lee J, Yoo SK, Kim K, et al. Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer. *Oncol Lett*. 2023;26(4):422. doi:10.3892/ol.2023.14008
36. Gupta A, Jain V, Singh A. Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications. *New Generation Computing*. 2022;40(4):987–1007. doi:10.1007/s00354-021-00144-0