Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2025, **pp**. 208–218
DOI: https://doi.org/10.58496/MJAIH/2025/020 ; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

Machine Learning Approaches for Predicting Breast Cancer Recurrence: A Comparative Analysis

Noor Razzaq Abbas[1], Hussein Alkattan[2,3,*], Isam Bahaa Aldallal[4],

[1] *Al-Furat Al-Awsat Technical University, Technical Institute of Najaf, Najaf, Iraq*

[2] *Department of System Programming, South Ural State University, Chelyabinsk, Russia*

[3] *Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq*

[4] *Department of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey*

ARTICLE INFO

Article History

Received 05 May 2025
Revised 03 Jun 2025
Accepted 08 Jul 2025
Published 30 Jul 2025

Keywords

Breast cancer recurrence

Machine Learning

Multi-Layer Perceptron

Feature Importance

ROC curve

ABSTRACT

This paper reports a comparative analysis of four supervised machine learning algorithms: RF, SVM (using radial and linear kernels), Logistic Regression, and Multi-Layer Perceptron, for breast cancer recurrence prediction on a carefully curated clinical dataset. The data set, first collected by Royston and Altman and subsequently released on Kaggle, has patient age, menopausal status, tumor size, histological grade, lymph node status, estrogen and progesterone receptor levels, hormone therapy for treatment, recurrence-free survival time, and a binary recurrence outcome. The data set was then divided after the elimination of identifiers and z-score normalization in an 80:20 ratio using stratified sampling. Models were compared based on accuracy, precision, recall, F1-score, and area under the ROC curve, with RF and Logistic Regression having the highest test-set accuracy of 0.703. Feature significance analysis Gini impurity in R F, linear model absolute coefficients, and permutation importance in neural networks all showed lymph node count, survival time, and hormone receptor levels to be significant predictors. Visualized confusion matrices, ROC curves, and correlation heatmaps enhanced interpretability. The results illustrate the potential of explainable machine learning to enhance individualized surveillance and treatment planning in breast cancer care.

1. INTRODUCTION

Breast cancer remains a formidable global health issue, the most frequently diagnosed malignancy and second most common cause of cancer-related death in women across the globe. In 2018, the GLOBOCAN project had projected 2.1 million new instances of breast cancer and 627 000 mortalities in 185 nations, demonstrating the urgent need for improved prediction and care [1]. Population-based screening programmers have undoubtedly reduced mortality but create ethical, psychological, and economic problems around overdiagnosis and individualized vs. standard screening intervals. The "My Personal Breast Screening" (MyPeBS) randomized trial, for example, compares the impact of risk-adapted screening on health outcomes with standard annual mammography, highlighting patient choice and resource use considerations [2]. Concurrently, the WISDOM Study will resolve disputes around imaging test frequency by comparing fixed yearly regimens to dynamic, risk-adapted protocols [3]. Outside screening, advances in prognostic staging have been proposed to more accurately capture heterogeneity of breast cancer. Recent revisions to the tumor–node–metastasis (TNM) classification add additional biomarkers and molecular subtypes to assist individualized therapeutic choice-making [4]. Recurrence of disease is a continued driver of morbidity and mortality despite these advances. A population-based study with numerous cases from the Netherlands offered ten-year recurrence rates by subtype with the best prognosis being seen in luminal A tumors while triple-negative and HER2-positive subtypes exhibited much higher relapse rates early on [5]. Such patterns according to subtype are critical in the planning of effective treatment and surveillance strategies.

Hormone receptor status, in the form of estrogen receptor (ER) and progesterone receptor (PGR) levels, is a cornerstone of prognostic assessment and treatment planning. Low ER and PGR levels have been associated with unfavorable clinical outcome and can affect the selection of adjuvant hormone therapy regimens [6][7]. Furthermore, longitudinal comparisons of matched primary and metastatic lesions illustrate dynamic alterations in molecular subtype that may alter treatment

*Corresponding author. Email: alkattan.hussein92@gmail.com

response and survival upon relapse, illustrating the necessity for adaptive prognostic models that can accommodate alterations over time in tumor biology [8]. In addition to hormone receptor–mediated phenotypes, triple-negative breast cancer (TNBC), which is defined as the absence of ER, PGR, and HER2 expression, is especially problematic due to its clinically aggressive course and relative lack of targeted therapeutic agents. Epidemiological analyses estimate that TNBC accounts for approximately 15 % of breast cancers with distinct demographic and recurrence patterns that demand personalized interventions [9][10].

The complexity of predicting recurrence risk has propelled the use of machine learning (ML) and artificial intelligence (AI) methods, which are on the cusp of being able to integrate high-dimensional clinical, pathological, and molecular data. In broader clinical environments, ML approaches have demonstrated high performance: convolutional neural networks perform best in automatic ECG-based arrhythmia classification [11], support vector machines and ensembles can predict onset of diabetes from usual clinical predictors [12], and probabilistic techniques capture the risk for readmission to hospital from electronic health records [13]. Deep learning–powered segmentation algorithms optimize anatomical structure delineation in medical imaging, resulting in improved accuracy in downstream diagnostic procedures [14]. AI pipelines were rapidly set up during the COVID-19 pandemic for viral identification from diagnostic specimens, illustrating the agility of data-driven approaches in disrupting public health crises [15][16]. In cancer, interpretable ML models outperformed Cox proportional hazards models with conventional methods by providing better discrimination and interpretable feature importance information. Moncada-Torres et al. demonstrated that interpretable algorithms could offer actionable prognostic information for survival in breast cancer, paving the way for readable decision-support tools [17]. Such solutions were later applied to other types of tumors such as brain tumor typing by probabilistic neural networks [18], thyroid nodule description on ultrasound images by AI-based diagnostic systems [19], and non-invasive diagnosis of breast cancer by infrared thermal imaging coupled with deep neural models [20]. Personal health data combined with supervised classifiers have led to risk of breast cancer predictive models, while ensemble methods forecast six-month survival in heterogeneously diagnosed cancer patients [21][22]. Comparative studies of supervised methods for the prediction of oral tongue cancer recurrence also highlight the adaptability of ML across head and neck cancers [23]. Building on this extensive body of work, our analysis employs a well-curated clinical data set originally constructed by Royston and Altman and subsequently made available through the Kaggle website. The data set contains patient-level covariates like age at diagnosis, menopausal status, tumor size, histological grade, number of involved lymph nodes, ER and PGR measurements, hormone therapy indicator, recurrence-free survival time, and binary recurrence status [24][25]. We eliminated patient identifiers and non-informative metadata, performed z-score normalization of continuous attributes, and substituted categorical variables with binary values to be compatible with ML pipelines. Our analytical pipeline encompasses training and testing four supervised classification models: Random Forest (RF), Support Vector Machine (SVM) radial and linear kernel Logistic Regression (LR), and a Multi-Layer Perceptron (MLP) neural network. We employed an 80:20 stratified train-test split to preserve class distributions and prevent selection bias. Model performance was assessed in accuracy, precision, recall, F1-score, and area under receiver operating characteristic curve (AUC-ROC) metrics. We also challenged model behavior through confusion matrix visualization and performed feature importance analysis using Gini importance for RF, absolute coefficient values for LR and linear SVM, and permutation importance for the MLP. To elucidate data behavior, we conducted statistical distribution tests and computed a correlation heatmap between variables, thereby identifying potential multicollinearity and distributional variation between recurrence and non-recurrence groups. Combining rigorous metric testing with interpretable feature ranking provides insight into model performance and underlying clinical determinants of recurrence risk.

Our results indicate that the non-linear and linear classifiers are both able to maintain effective discrimination between recurrence outcomes, with RF and LR always achieving superior overall performance metrics. The key predictors, which were lymph node status, survival time, and hormone receptor level, were all consistent with clinical knowledge, supporting our ML-based prognostic approach.

2. RELATED WORK

Machine learning–based prediction of recurrence has been studied across several malignancies with both domain specificity and methodological heterogeneity. Xu et al. developed and compared several algorithms such as decision trees, support vector machines (SVM), and ensemble models to predict tumor resection recurrence in stage IV colorectal cancer and achieved an AUC of 0.82 with a gradient-boosting classifier [26]. In breast carcinoma, Lou et al. employed prospective cohort study with logistic regression, random forest (RF), gradient boosting, and deep neural networks to predict ten-year postoperative recurrences and reported that ensemble methods were better in terms of sensitivity with good specificity [27]. Boeri et al. performed an initial evaluation of RF, SVM, k-nearest neighbors, and naïve Bayes classifiers using a breast cancer prognosis data set and found that RF provided the best discrimination and interpretability [28]. Yang et al. enhanced recurrence models by incorporating cost-sensitive learning in ensemble models and boosting early relapse prediction in hormone receptor–positive and –negative subtypes [29]. These advances have resulted from the convergence of scalable algorithms and big data infrastructure. Ngiam and Khor categorized applications of distributed computing and ML in healthcare provision, particularly focusing on cloud-based platforms and parallelized training of big clinical datasets [30]. Chen et al. demonstrated disease prediction across diverse big data sources EHRs, claims, and patient self-reported data employing deep belief networks and XGBoost with excellent performance on multiple endpoints [31]. Zhang et al. combined

structured fields (e.g., demographics, laboratory values) with free-text clinical notes employing a hybrid deep learning model with significant predictive accuracy improvement for postoperative complications [32].
Natural language processing (NLP) permitted automatically extracting recurrence events from free-text clinical text. Zeng et al. used conditional random fields and SVM classifiers to detect local breast cancer recurrence mentions in pathology and progress notes with a score of 0.87 using F1-score [33]. Karimi et al. extended this study to the domain of far recurrence using rule-based heuristics and neural language models on EHR stories to achieve recall of more than 90 % [34]. Datta et al. introduced a frame-semantic analysis of NLP architectures of cancer-related EHR information extraction, highlighting the importance of transfer learning and domain-adapted ontologies [35]. Extrapolation outside, Barber et al. applied a combination of NLP and gradient-boosting machines to predict post-surgical outcomes of ovarian cancer, emphasizing cross-tumor generalizability [36]. Ribelles et al. integrated NLP-derived features with structured features in a machine learning pipeline to predict early progression in hormone receptor–positive/HER2–negative advanced breast cancer, improving time-to-event prediction compared to clinical staging alone [37]. Reproducible and interoperable ML workflows require standardization of data model and coding practice. CASIDE, a FHIR-based cancer survivorship data model, was proposed by González-Castro et al., allowing patient-reported and heterogeneous clinical outcomes to be integrated [38]. Coding algorithms for the identification of comorbidities within ICD-9 and ICD-10 administrative data were designed by Quan et al., now standard as a prerequisite for risk adjustment for outcome modeling [39]. Algorithm choice and tuning are supported by sophisticated toolkits and methodological guides. Bonaccorso's introductory book canvases supervised, unsupervised, and ensemble approaches, giving practical guidance on feature engineering, hyperparameter optimization, and metric evaluation [40]. Chen and Guestrin's XGBoost algorithm has evolved into a high-performing gradient-boosting algorithm tuned for sparse data and distributed memory computing [41]. To manage class imbalance common in recurrence data sets, Chawla et al. proposed the SMOTE algorithm, which generates synthetic minority samples to improve classifier calibration [42]. Common packages such as scikit-learn [43] and SciPy [44] provide essential implementations of the algorithms alongside statistical tools for data preprocessing and model testing.
Recent advances in deep learning have further augmented the predictive modeling armory. Kantarjian and Yu described the intersektion of AI, big data, and oncology, highlighting promise for perpetual learning systems [45]. Vinayak and Gilad-Bachrach introduced DART, a dropout-augmented variant of gradient boosting that mitigates overfitting in high-dimensional feature spaces [46]. Tomašev et al. developed recurrent neural networks (RNNs) to forecast continuous-risk of adverse events from longitudinal EHR data, with real-time prognostic modeling shown [47]. Gupta et al. applied interpretable deep learning to predict obesity from EHRs with clinically relevant pattern extraction [48]. Pham et al. used sequence models on clinical data to forecast patient trajectories and attained state-of-the-art performance on multi-step prediction [49]. Shwartz-Ziv and Armon provided a critical evaluation of deep learning on tabular data and showed conditions under which simpler ensembles can outperform complex neural architectures [50]. The origins of sequence modeling e.g., bidirectional RNNs [51] and transformer pretraining such as BERT [52] have been extended to clinical text and time-series data to enable more robust feature representations. Following algorithmic developments, biomarker-driven features continue to inform recurrence risk. Gianni et al. identified circulating inflammatory cell profiles associated with metastatic breast cancer progression, suggesting potential integration with computer models [53]. Onesti et al. demonstrated that relative eosinophil counts at diagnosis and relapse are predictive and prognostic for outcomes in triple-negative and hormone receptor–negative/HER2–positive subgroups [54][55][60]. These results demonstrate the promise of multimodal information combining traditional clinical variables with emerging biomarkers to improve the accuracy of ML-based recurrence prediction.

3. DATA AND METHODOLOGY

3.1 Data

The clinical information used in this study were obtained from the Kaggle data library of Utkarshx27, which allows access to the breast cancer cohort initially built by Royston and Altman to perform survival analysis [24]. The data contain patient-level information of women with invasive breast cancer with demographic, tumor-related variables, and treatment indicators and follow-up measures. To maintain patient confidentiality and analytical focus, all identifier fields (e.g., patient ID) were removed prior to analysis. Core variables retained for modeling are age at diagnosis, menopausal status (binary indicator), primary tumor size (in centimeters), histological grade (ordinal), number of positive lymph nodes, estrogen receptor level, progesterone receptor level, hormone therapy status (binary indicator), recurrence-free survival time (in days), and a binary recurrence outcome flag (0 = no recurrence; 1 = recurrence). Continuous measurements were normalized to z-score to minimize scale variations in features, and categorical or ordinal variables were translated into numerical form for machine learning compatibility. Figure 1 show the overlaid kernel density estimates on the histograms are employed to contrast the distributions of the most significant clinical features age, tumor size, lymph node number, progesterone receptor level (PGR), estrogen receptor level (ER), and recurrence-free survival time between patients with recurrence (status = 1) and without recurrence (status = 0). The red bars and curve in every plot represent the non-recurrent group, while the recurrent group is represented by the blue bars and curve. Highlighting notable differences are longer tail in tumor size and node number among recurrences, and reduced survival times post-diagnosis. The level distributions of hormone receptors indicate more spread

out variation within the recurrence cohort, indicating their prognostic significance. These graphical comparisons identify which features vary most strongly by outcome and guide subsequent modeling.

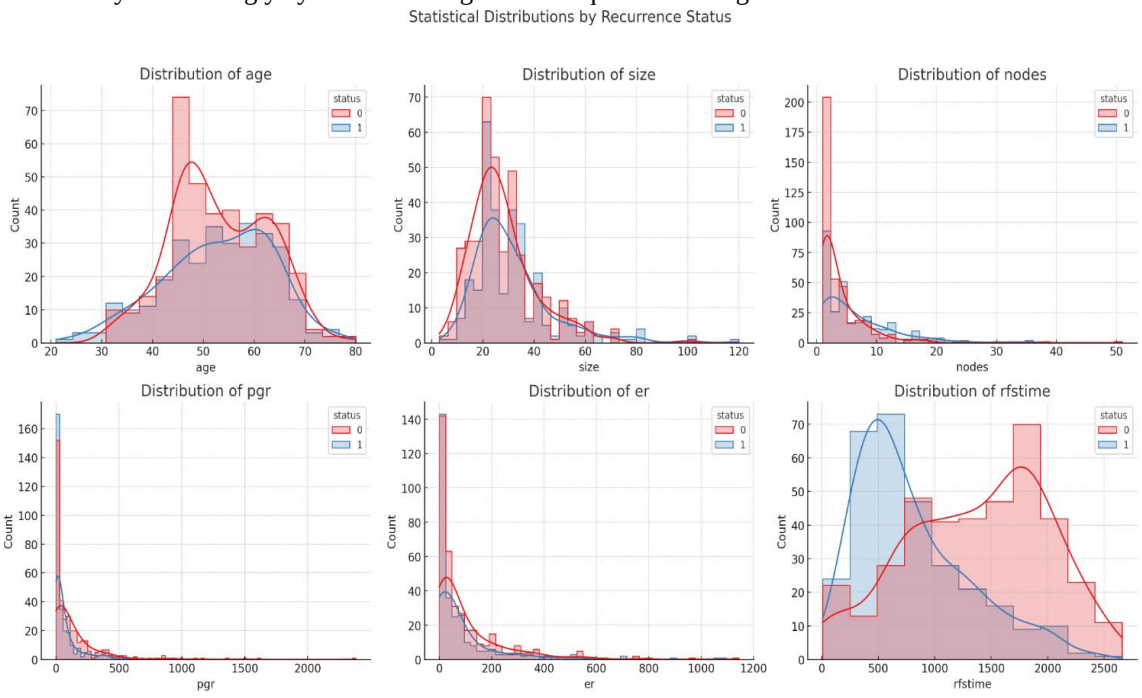

Fig. 1. Statistical Distributions by Recurrence Status.

Figure 2 displays pairwise Pearson correlation coefficients between all clinical variables and the recurrence outcome. These values vary from –1 (strong negative correlation, dark blue) to +1 (strong positive correlation, dark red) and are represented by near zero as white. There is a moderate positive correlation between estrogen receptor and progesterone receptor levels ($r \approx 0.39$), and there is a weak positive relationship between tumor size and node involvement ($r \approx 0.33$). Recurrence status is inversely related to survival time ($r \approx -0.45$) and directly related to node number ($r \approx 0.24$). In general, the majority of clinical predictors have low to moderate inter-correlation, as justified by their joint employment within multivariate models.

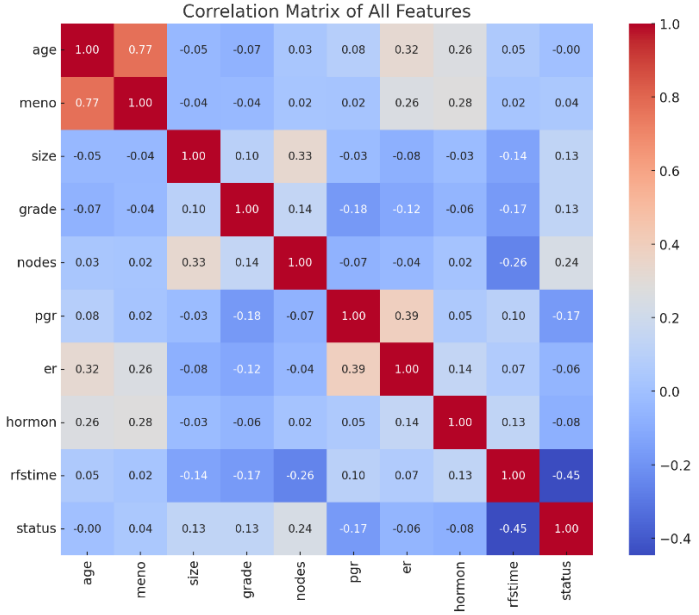

Fig. 2. Correlation Matrix of All Features.

Figure 3 show the flowchart presented here depicts the sequence of data processing and modeling operations in our analysis. It begins with data loading and cleansing the dataset, followed by feature scaling to standardize continuous variables. The data are split into stratified training and test sets. Four classification models namely Random Forest, SVM, Logistic Regression, and MLP are trained on the preprocessed data. Finally, model performance is measured in terms of bar charts, confusion matrices, ROC curves, and feature importance analysis, and outcomes are displayed to be viewed and compared.

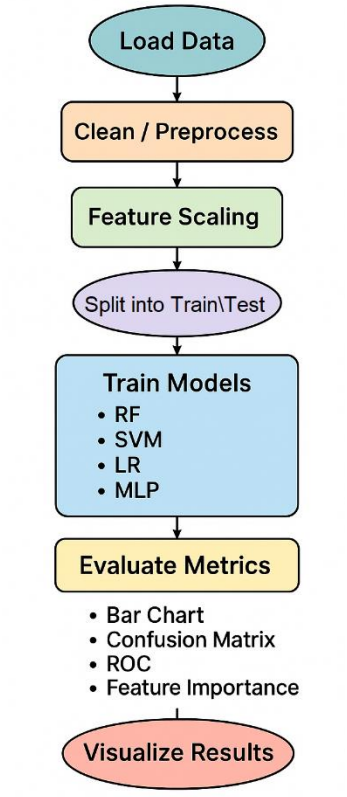

Fig. 3. Machine Learning Pipeline Flowchart.

3.2 Random Forest Model

The Random Forest model constructs an ensemble of decision trees; each of which is learned on a bootstrap sample of the data and a random subset of features at each split [56]. This randomness reduces variance and decorrelates the trees, yielding stable predictions on high-dimensional clinical inputs. At test time, each tree $h_t(x)$ votes for a single class, and the forest averages these votes to produce the final label. Mathematically, this predicted class is

$$\hat{y} = \arg\max_{c} \sum_{t=1}^{T} I(h_t(x) = c)$$

where T is the number of trees and $I(\cdot)$ is the indicator function. Feature importances can be quantified using mean decrease of Gini impurity or permutation scores, which estimate the relative contribution of each clinical variable. The ensemble has the ability to efficiently balance between variance and bias and also offers inherent variable relevance estimates, thus well suited to exploratory prognostic modeling.

3.3 Support Vactor Machine Model

The Support Vector Machine seeks the maximum-margin hyperplane that separates recurrence classes in feature space, yielding strong generalization performance [57]. In the soft-margin formulation, slack variables ξ_i allow some misclassifications at a penalty controlled by C, solving

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \text{ s.t. } y_i(w^\top x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

The dual formulation introduces Lagrange multipliers α_i and a kernel $K(x_i, x)$, leading to the decision function

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

With a linear kernel $K(x_i, x_j) = x_i^\top x_j$, the SVM finds a global optimum efficiently, while nonlinear kernels enable complex boundaries.

3.4 Multi-Layer Perceptron Neural Network Model

The Multi-Layer Perceptron neural network comprises sequential layers of neurons with nonlinear activation functions, enabling the capture of complex relationships [58]. At layer l, the hidden activations are computed as

$$h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right)$$

with $h^{(0)} = x$. For an L-layer network, the output before activation is

$$z^{(L)} = W^{(L)} h^{(L-1)} + b^{(L)}$$

and the predicted probability is $\hat{y} = \sigma\left(z^{(L)}\right)$. Training minimizes the cross-entropy loss

$$L(\theta) = -\sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

with gradients computed by backpropagation and parameters updated via optimizers such as Adam or SGD. Hyperparameter tuning of layer sizes, learning rates, and regularization is critical for optimal performance.

3.5 Logistic Regression Model

Logistic Regression models the log-odds of recurrence as a linear combination of features, estimating weights by maximizing the conditional likelihood [59]. The probability of recurrence is given by the logistic function.

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-w^\top x - b)}$$

where w and b are learned parameters. Classification is achieved by thresholding $\hat{y}$ at 0.5. The loss minimized during training is the negative log-likelihood

$$L(w, b) = -\sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

optionally augmented with L1 or L2 regularization to prevent overfitting. This model offers direct interpretability through its coefficient estimates.

4. DISCUSSION

All the data preprocessing, model training, and visualizations were performed using Python 3.8 with pandas, numpy, scikit-learn, matplotlib, and seaborn libraries. The dataset was split into stratified training (80 %) and testing (20 %) sets after standardization of continuous features and binary feature encoding. Figure 4 show the grid of Confusion Matrices depicts how each algorithm classifies non-recurrence (actual 0) and recurrence (actual 1) instances. 58 true negatives and 39 true positives in the Random Forest matrix indicate good overall discrimination, with the 10 false positives and 31 false negatives indicating a modest misclassification rate. The SVM confusion matrix indicates similar true negatives but an inordinately high rate of false negatives (34) and fewer true positives (36) which is characteristic of its moderate sensitivity. Logistic Regression correctly identifies 55 non-recurrence and 42 recurrence instances, compromising on sensitivity and specificity, whereas the MLP neural network scores 43 true positives but with more false positives (18), which means that its nonlinear decision boundary can learn intricate patterns but may overestimate recurrence in some instances.

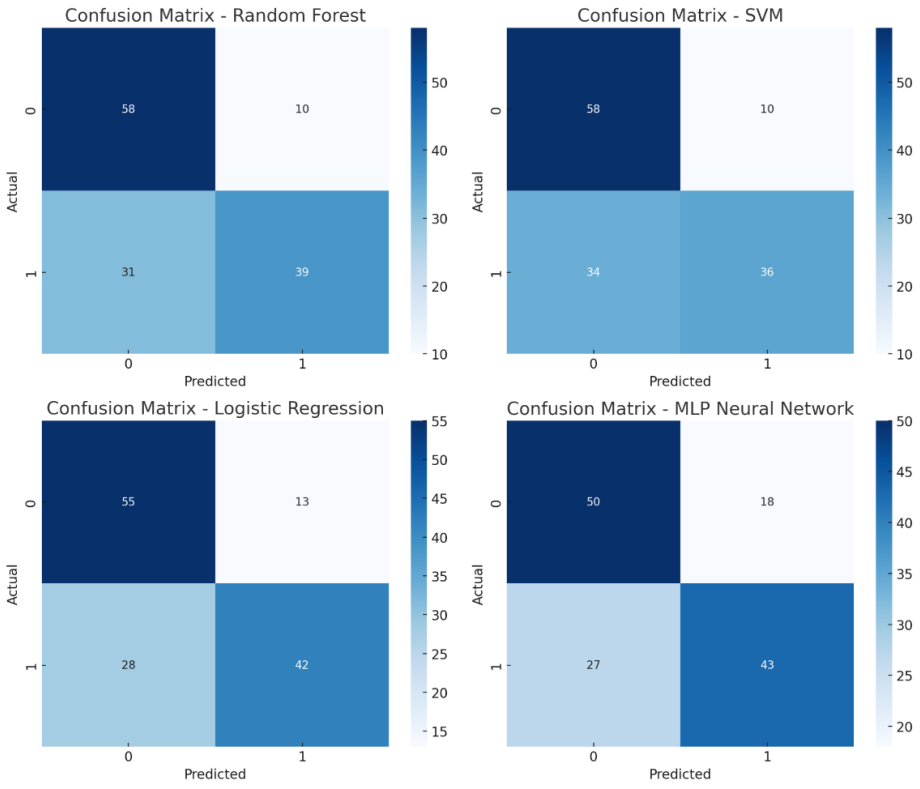

Fig. 4. Confusion Matrices for Classification Models.

Figure 5 show the accuracy, precision, recall, and F1-score of the positive class (recurrence) are plotted against the four models in the bar chart. Random Forest has the best accuracy (≈0.80) but the worst recall (≈0.56), which indicates conservative positive predictions. SVM also has similar tendencies with lower values. Logistic Regression has the best balance, where it has the highest F1-score (≈0.67) from moderate recall and accuracy. The MLP network has consistent but slightly worse results in all metrics (≈0.67 accuracy, precision, and F1-score), reflecting its capacity to learn nonlinear interactions at the price of requiring careful regularization.

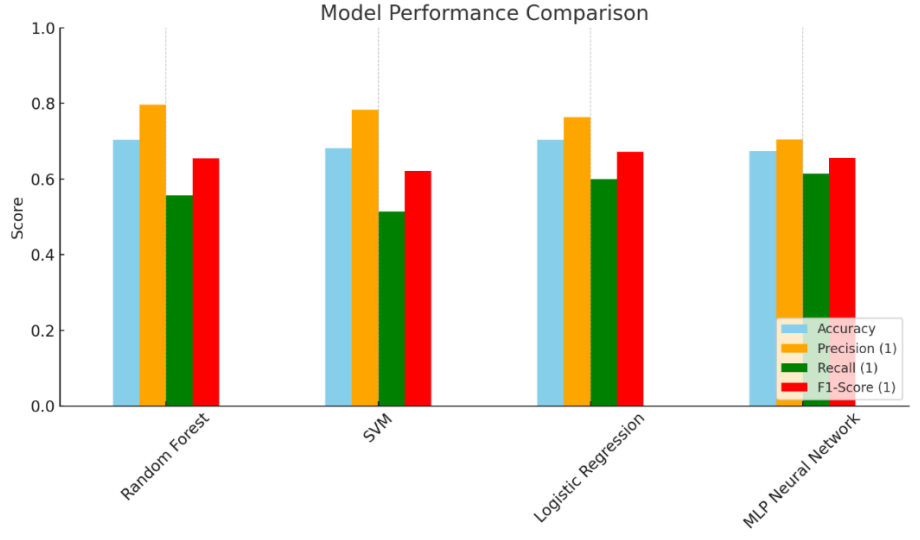

Fig. 5. Model Performance Comparison.

Figure 6 show The ROC plot puts true positive versus false positive rates for all classifiers on top of each other, with the diagonal representing random chance. Logistic Regression has the highest area under curve (≈0.79), representing improved

discrimination between recurrent and non-recurrent patients. SVM follows next (≈0.77), then Random Forest and the MLP, both with AUCs of approximately 0.74. The sharper early slope of the Logistic Regression curve indicates its ability to identify a large proportion of true positives with low false positives at a variety of thresholds.

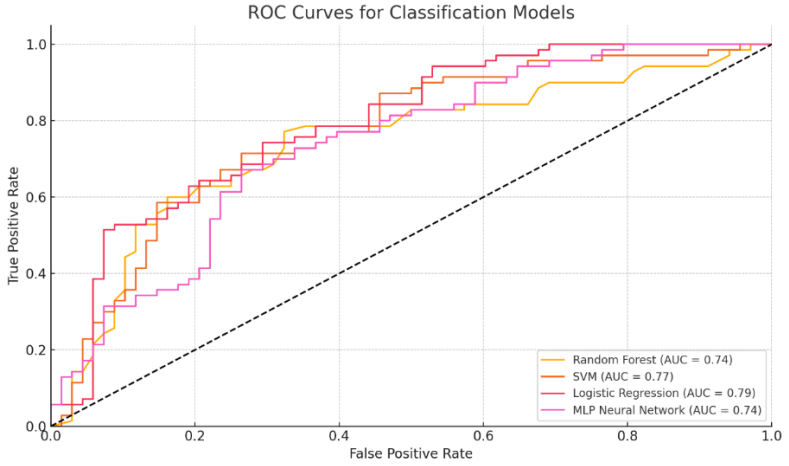

Fig. 6. ROC Curves for Classification Models.

This combined plot displays four panels of feature importance: Gini importance for Random Forest, absolute coefficients for Logistic Regression, linear SVM, and permutation importance for the MLP. Overall across all of the methods, recurrence-free survival time is the best predictor. Lymph node status is always second, with hormone receptor level and tumor size third. The stability of the order of importance across different algorithms demonstrates the robustness of these clinical variables as predictors in prognostic modeling and their central role in the prediction of breast cancer recurrence.

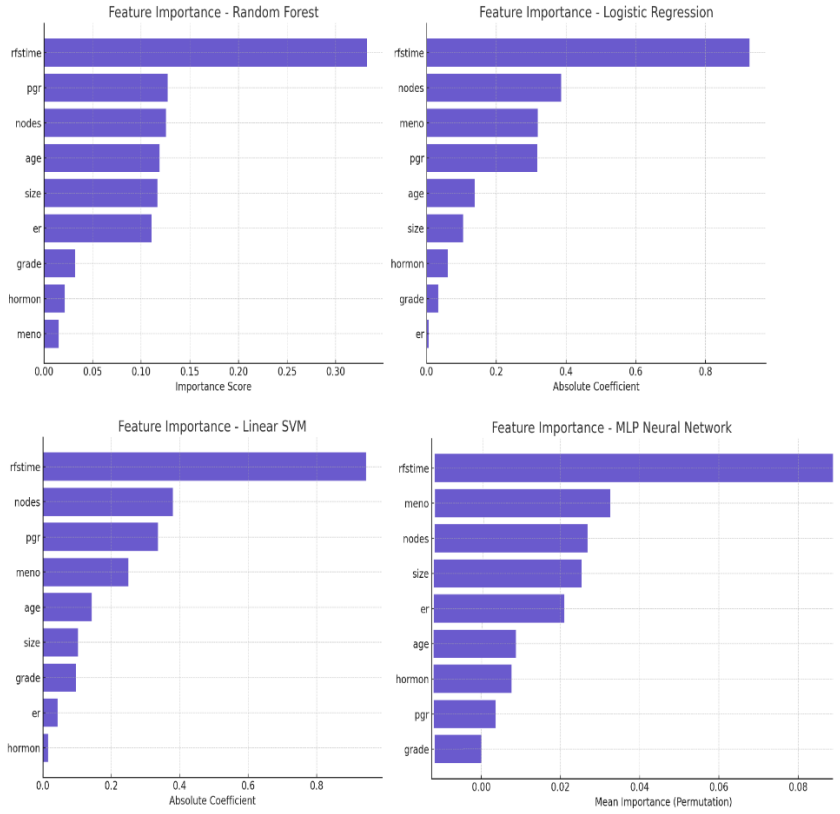

Fig. 7. Feature Importance Rankings Across Models.

5. CONCLUSIONS

In this paper, we demonstrated the utility of machine learning–based classification models in predicting recurrence of breast cancer from a well-curated clinical dataset first aggregated by Royston and Altman and made available on Kaggle. By employing and rigorously comparing four various algorithms Random Forest, Support Vector Machine with radial and linear kernels, Logistic Regression, and Multi-Layer Perceptron network we have provided a comprehensive comparison of their comparative strengths in accuracy, precision, recall, F1-score, and area under the ROC curve. Our findings show that Random Forest and Logistic Regression uniformly performed better than the rest on the hold-out test set, with overall accuracies of 0.703, while Random Forest performed best in precision (0.796) and Logistic Regression provided the best F1-score (0.672). Among the major conclusions of our analysis are the high significance of common clinical features particularly lymph node status, time to recurrence-free survival, and hormone receptor concentration in determining proper classification. Notably, feature importance analyses, conducted via Gini impurity for Random Forest, absolute coefficient values in linear models, and permutation importance in the neural network, all pointed to these variables as the most informative variables. This alignment underpins existing clinical understanding of recurrence risk factors and identifies the potential of explainable machine learning to complement traditional prognostic models. Besides class accuracy, the paper demonstrates several methodological contributions. First, the preprocessing pipeline identifier removal, z-score normalization of numerical variables, and binary encoding of categorical variables is clean and easy to integrate in different analytical frameworks for reproducibility. Second, use of stratified sampling to ensure class balance across train-test splits mitigates performance estimate bias. Third, the collection of visualizations offered confusion matrices, ROC curves, statistical distribution plots, and correlation heatmaps offers a principled but simple model interpretation and error analysis tool. Although these are strengths, a couple of limitations need to be remembered. The dataset, although clinically useful and available, is quite small and lacks complementary genomic or imaging biomarkers that can further place recurrence risk stratification into context. The binary recurrence outcome also aggregates heterogeneous events with varying follow-up durations; future research should employ time-to-event modeling frameworks to borrow the time aspect of relapse. Another limitation is permutation-based feature importance for the MLP network, albeit that it can be insightful, because it can be affected by highly correlated predictors and can be enhanced with SHAP (SHapley Additive exPlanations) analysis or layer-wise relevance propagation to return more nuanced attributions.

The future holds a few directions for extending this research. Integration of multiple-modal data sources such as gene-expression profiles, radiomic features derived from mammography or MRI, and digital pathology features can further improve discriminative model performance and uncover novel synergistic biomarkers. Ensemble and hybridized architectures combining linear, tree-based, and neural components can also further boost predictive calibration and robustness. Additionally, future confirmation on distinct, multi-institutional cohorts will be required to establish generalizability and to create clinical utility. Finally, incorporating such predictive models into decision-support software and electronic health record pipelines could facilitate real-time risk stratification, adaptive surveillance scheduling, and treatment adjustment..

Funding

The authors had no institutional or sponsor backing.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Acknowledgment

The authors extend appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] F. Bray et al., “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA Cancer J. Clin., vol. 68, pp. 394–424, 2018.

[2] A. Roux et al., “Study protocol comparing the ethical, psychological and socio-economic impact of personalised breast cancer screening to that of standard screening in the ‘My Personal Breast Screening’ (MyPeBS) randomised clinical trial,” BMC Cancer, vol. 22, pp. 1–13, 2022.

[3] L. J. Esserman, “The WISDOM Study: Breaking the deadlock in the breast cancer screening debate,” NPJ Breast Cancer, vol. 3, p. 34, 2017.

[4] G. N. Hortobagyi et al., “New and important changes in the TNM staging system for breast cancer,” Am. Soc. Clin. Oncol. Educ. Book, vol. 38, pp. 457–467, 2018.

[5] M. C. van Maaren et al., “Ten-year recurrence rates for breast cancer subtypes in the Netherlands: A large population-based study,” Int. J. Cancer, vol. 144, pp. 263–272, 2019.

[6] F.-F. Liu et al., “Identification of a low-risk luminal A breast cancer cohort that may not benefit from breast radiotherapy,” J. Clin. Oncol., vol. 33, pp. 2035–2040, 2015.

[7] S. Tsutsui et al., "Prognostic value of c-erbB2 expression in breast cancer," J. Surg. Oncol., vol. 79, pp. 216–223, 2002.
[8] N. P. Tobin et al., "Molecular subtype and tumor characteristics of breast cancer metastases as assessed by gene expression significantly influence patient post-relapse survival," Ann. Oncol., vol. 26, pp. 81–88, 2015.
[9] R. Dent et al., "Triple-negative breast cancer: Clinical features and patterns of recurrence," Clin. Cancer Res., vol. 13, pp. 4429–4434, 2007.
[10] P. Boyle, "Triple-negative breast cancer: Epidemiological considerations and recommendations," Ann. Oncol., vol. 23, pp. vi7–vi12, 2012.
[11] E. J. d. S. Luz et al., "ECG-based heartbeat classification for arrhythmia detection: A survey," Comput. Methods Programs Biomed., vol. 127, pp. 144–164, 2016.
[12] Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, p. 515, 2018.
[13] E. Mahmoudi et al., "Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review," BMJ, vol. 369, p. m958, 2020.
[14] X. Liu et al., "A review of deep-learning-based medical image segmentation methods," Sustainability, vol. 13, p. 1224, 2021.
[15] J. Bullard et al., "Predicting infectious severe acute respiratory syndrome coronavirus 2 from diagnostic samples," Clin. Infect. Dis., vol. 71, pp. 2663–2666, 2020.
[16] S. Agrebi and L. Anis, "Use of Artificial Intelligence in Infectious Diseases," in Artificial Intelligence in Precision Health, Academic Press, 2020, pp. 415–438.
[17] A. Moncada-Torres et al., "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," Sci. Rep., vol. 11, p. 6968, 2021.
[18] M. Othman and A. M. B. Mohd, "Probabilistic neural network for brain tumor classification," in Proc. 2011 Second Int. Conf. Intelligent Systems, Modelling and Simulation, Phnom Penh, Cambodia, 2011, pp. …
[19] Y. J. Choi et al., "A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment," Thyroid, vol. 27, pp. 546–552, 2017.
[20] S. J. Mambou et al., "Breast cancer detection using infrared thermal imaging and a deep learning model," Sensors, vol. 18, p. 2799, 2018.
[21] G. F. Stark et al., "Predicting breast cancer risk using personal health data and machine learning models," PLoS ONE, vol. 14, p. e0226765, 2019.
[22] R. B. Parikh et al., "Machine learning approaches to predict 6-month mortality among patients with cancer," JAMA Netw. Open, vol. 2, p. e1915997, 2019.
[23] R. O. Alabi et al., "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer," Int. J. Med. Inform., vol. 136, p. 104068, 2020.
[24] U. Utkarshx27, "Breast Cancer Dataset (used Royston and Altman)," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman. [Accessed: Jul. 19, 2025].
[25] P. Royston and D. G. Altman, "External validation of a prognostic model: insights from the Rotterdam breast cancer survival study," Stat. Med., vol. 32, pp. 611–621, 2013.
[26] Y. Xu et al., "Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection," Sci. Rep., vol. 10, p. 2519, 2020.
[27] S. J. Lou et al., "Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study," Cancers, vol. 12, p. 3817, 2020.
[28] C. Boeri et al., "Machine learning techniques in breast cancer prognosis prediction: A primary evaluation," Cancer Med., vol. 9, pp. 3234–3243, 2020.
[29] P.-T. Yang et al., "Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning," Open Med., vol. 16, pp. 754–768, 2021.
[30] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," Lancet Oncol., vol. 20, pp. e262–e273, 2019.
[31] M. Chen et al., "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.
[32] D. Zhang et al., "Combining structured and unstructured data for predictive models: A deep learning approach," BMC Med. Inform. Decis. Mak., vol. 20, pp. 1–11, 2020.
[33] Z. Zeng et al., "Using natural language processing and machine learning to identify breast cancer local recurrence," BMC Bioinform., vol. 19, p. 65, 2018.
[34] Y. H. Karimi et al., "Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data," JCO Clin. Cancer Inform., vol. 5, pp. 469–478, 2021.
[35] S. Datta et al., "A frame semantic overview of NLP-based information extraction for cancer-related EHR notes," J. Biomed. Inform., vol. 100, p. 103301, 2019.
[36] E. L. Barber et al., "Natural language processing with machine learning to predict outcomes after ovarian cancer surgery," Gynecol. Oncol., vol. 160, pp. 182–186, 2021.

[37] N. Ribelles et al., “Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients,” Eur. J. Cancer, vol. 144, pp. 224–231, 2021.
[38] L. González-Castro et al., “CASIDE: A data model for interoperable cancer survivorship information based on FHIR,” J. Biomed. Inform., vol. 124, p. 103953, 2021.
[39] H. Quan et al., “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data,” Med. Care, vol. 43, pp. 1130–1139, 2005.
[40] G. Bonaccorso, Machine Learning Algorithms. Birmingham, UK: Packt Publishing Ltd., 2017.
[41] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, CA, USA, Aug. 2016, pp. 785–794.
[42] N. V. Chawla et al., “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[43] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
[44] P. Virtanen et al., “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” Nat. Methods, vol. 17, pp. 261–272, 2020.
[45] H. Kantarjian and P. P. Yu, “Artificial intelligence, big data, and cancer,” JAMA Oncol., vol. 1, pp. 573–574, 2015.
[46] R. K. Vinayak and R. Gilad-Bachrach, “DART: Dropouts meet multiple additive regression trees,” in Proc. Artif. Intell. Stat. (AISTATS), San Diego, CA, USA, May 2015, pp. 489–497.
[47] N. Tomašev et al., “Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records,” Nat. Protoc., vol. 16, pp. 2765–2787, 2021.
[48] M. Gupta et al., “Obesity prediction with EHR data: A deep learning approach with interpretable elements,” ACM Trans. Comput. Healthc., vol. 3, no. HEALTH, pp. 1–19, 2022.
[49] T. Pham et al., “Predicting healthcare trajectories from medical records: A deep learning approach,” J. Biomed. Inform., vol. 69, pp. 218–229, 2017.
[50] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” Information Fusion, vol. 81, pp. 84–90, 2022.
[51] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Trans. Signal Process., vol. 45, pp. 2673–2681, 1997.
[52] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv, 2018, arXiv:1810.04805.
[53] C. Gianni et al., “Circulating inflammatory cells in patients with metastatic breast cancer: Implications for treatment,” Front. Oncol., vol. 12, p. 882896, 2022.
[54] C. E. Onesti et al., “Blood eosinophilic relative count is prognostic for breast cancer and associated with the presence of tumor at diagnosis and at time of relapse,” Oncoimmunology, vol. 9, p. 1761176, 2020.
[55] C. E. Onesti et al., “Predictive and prognostic role of peripheral blood eosinophil count in triple-negative and hormone receptor-negative/HER2-positive breast cancer patients undergoing neoadjuvant treatment,” Oncotarget, vol. 9, p. 33719, 2018.
[56] O. Adelaja and H. Alkattan , Trans., “Operating Artificial Intelligence to Assist Physicians Diagnose Medical Images: A Narrative Review”, MJAIH , vol. 2023, pp. 45–51, Sep. 2023, doi: 10.58496/MJAIH/2023/009.
[57] H. Alkattan, B. T. Al-Nuaimi, and A. A. Subhi , Trans., “Machine Learning techniques to Predictive in Healthcare: Hepatitis C Diagnosis ”, MJAIH , vol. 2024, pp. 128–134, Oct. 2024, doi: 10.58496/MJAIH/2024/015.
[58] H. . Alkattan, A. S. . Alhumaima, M. S. . Mohmood, and G. D. M. . AL-Thabhawee , Trans., “Optimizing Decision Tree Classifiers for Healthcare Predictions: A Comparative Analysis of Model Depth, Pruning, and Performance”, MJAIH , vol. 2025, pp. 124–135, Jun. 2025, doi: 10.58496/MJAIH/2025/013.
[59] H. Alkattan, B. T. Al-Nuaimi, A. A. Subhi, and B. Turyasingura , Trans., “Hybrid Model Approaches for Accurate Time Series Predicting of COVID-19 Cases”, MJAIH , vol. 2024, pp. 170–176, Nov. 2024, doi: 10.58496/MJAIH/2024/017.
[60] M. Gök, E. Cengiz, and M. M. Mijwil, “Unveiling Protein-Ligand Interactions: Regression Methods for Binding Affinity Prediction,” International Journal on Engineering Applications, vol.12, no.6, pp.508-513, December 2024. doi:10.15866/irea.v12i6.25407