DOI: 10.37943/24UTRW4400

Zamart Ramazanova
MS in Physics, Researcher, Department of Electrical and Computer Engineering and National Laboratory Astana
zamart.ramazanova@nu.edu.kz, orcid.org/0000-0002-2623-8901
Nazarbayev University, Kazakhstan
Yeldar Baiken
Ph.D., Researcher, National Laboratory Astana
yebaiken@nu.edu.kz, orcid.org/0000-0003-1742-2536
Nazarbayev University, Kazakhstan
Senior researcher, Center for BioEnergy Research LLP
Bakhyt Matkarimov
Dr.Sci., Leading Researcher, National Laboratory Astana
bmatkarimov@nu.edu.kz, orcid.org/0000-0003-0775-7324
Nazarbayev University, Kazakhstan
Arshat Urazbayev
Ph.D., Senior Researcher, National Laboratory Astana
arshat@gmail.com, orcid.org/0000-0002-4763-1438
Nazarbayev University, Kazakhstan
Askhat Myngbay
Ph.D., Senior Researcher, National Laboratory Astana
askhat.myngbay@yahoo.com, orcid.org/0000-0002-3867-847X
K.Zhubanov Aktobe Regional University, Kazakhstan
Bauyrzhan Aituov
General Director
aituov@cber.kz, orcid.org/0009-0008-3001-8144
Center for BioEnergy Research LLP, Kazakhstan

AN INFORMATION TECHNOLOGY APPROACH TO PREDICT BREAST CANCER USING MACHINE LEARNING

Abstract: Breast cancer continues to be the most encountered malignancy in women globally and a leading cause of cancer-related mortality. This study describes an Information Technology approach to evaluate interpretable machine-learning methods for breast cancer prediction using routine clinical data and to situate performance against prior literature. All calculations are based on the Breast Cancer Wisconsin Diagnostic dataset (569 instances; malignant/benign labels) hosted by the UCI Machine Learning Repository. Each sample corresponds to a breast mass classified as malignant or benign. Four supervised machine learning models were applied: Logistic Regression with L1 penalty, Random Forest, Decision Tree, and Naïve Bayes, and compared the area under the ROC curve (AUC), accuracy, sensitivity, and specificity using DeLong's test with Holm correction. The reproducible pipeline consisted of preprocessing, recursive feature elimination for feature selection, and a 5-fold cross-validation for hyperparameter tuning. Among the four models, the L1-penalized Logistic Regression yielded the best results, with an AUC indicating accuracy, sensitivity, and specificity of 99.6% (97.3%, 95.2%, 98.6%) on the test sets, respectively. This study illustrates the effective integration of supervised machine learning methods into diagnostic systems to produce early, accurate, interpretable diagnoses of disease. This study reinforces the proposed information technology approach for breast cancer prognosis. Limitations of the study are a moderately sized, homogeneous cohort, and restricted focus on structured variables, which may enhance internal validity while restricting generalizability. Our findings contribute to an emerging body of literature that well-tuned, regularized logistic regression provides a reasonable baseline against which breast cancer risk and other study outcomes can be compared, and a pragmatic route toward trustworthy AI in oncology.

Keywords: information technology; breast cancer; machine learning; model and feature selection; 5-fold cross-validation.

Introduction

Breast cancer remains one of the leading causes of cancer-related morbidity and mortality among women worldwide, with early detection playing a critical role in improving survival rates [1]. In recent years, the active development of information technology (IT) and the emergence of accessible computing resources have facilitated the widespread use of machine learning methods in medicine. These methods have become especially promising for analyzing large datasets of clinical and biomedical data to create effective prognostic models [2]. These technologies offer new opportunities for analyzing large volumes of clinical and biomedical data, facilitating the development of robust predictive models. Early detection of breast cancer is crucial because it significantly improves the chances of successful treatment. Yet, traditional diagnostic methods like mammography, histopathological analysis, and genetic testing face challenges such as variability in interpretation, high costs, and the need for specialized expertise [1], [3]. Machine learning (ML), supported by modern information technology, offers a promising solution by quickly analyzing large sets of clinical and biological data to find meaningful patterns [2], [4]. This approach makes diagnoses more accurate and quicker, allows for tailored treatments, eases pressure on healthcare systems, and helps doctors make better, faster decisions.

Recently, artificial intelligence (AI) and ML have improved our ability to distinguish malignant tumors from benign ones with high sensitivity and specificity. Techniques such as support vector machines (SVM), random forests, and convolutional neural networks (CNN) have shown success in medical imaging, histopathological analysis, and molecular data [5]. However, significant challenges exist with variability in the data used, lack of consistent standards, and limited interpretability produced by machine learning methods [6]. Many research studies remain limited to using a single data type. It limits the generalizability of their results to more diverse patient populations [7]. For example, models based solely on mammographic images may not account for features of tumor heterogeneity revealed by histopathological or genomic data. Conversely, approaches centered on genomic information may overlook important structural or morphological features captured through imaging [8], [9].

Our research aims to develop a robust approach that integrates multiple medical data sources to improve breast cancer detection accuracy and clinical utility. A multimodal method is important in precision medicine, where individualized risk assessment and targeted therapies improve patient outcomes.

Recent research, including Lu et al. [10], demonstrates that integrating histopathological imagery with genomic data enhances the accuracy of cancer subtype classification. Similarly, other investigations have effectively combined radiomics and transcriptomics to facilitate early cancer detection [11]. These cases underscore the importance of synthesizing varied data sources while prioritizing model interpretability and clinical relevance.

We propose that a multimodal, ML-based predictive model can predict breast cancer more accurately than traditional diagnostic methods alone. To test this, our methodology includes preparing the data, selecting important features, training models, and evaluating their performance carefully. Specifically, we will: 1) test four supervised ML models L1-regularized Logistic Regression, Random Forest, Decision Tree, and Naïve Bayes using the Breast Cancer Wisconsin Diagnostic dataset; 2) use recursive feature elimination (RFE) to select the best diagnostic features; 3) compare model performance through cross-validation and testing, measuring accuracy, sensitivity, specificity, and AUC; and 4) demonstrate the value of interpretable ML models in clinical environments.

Methods and Materials

Data information

All calculations are based on the Breast Cancer Wisconsin (Diagnostic) dataset curated by Wolberg et al. and hosted by the UCI Machine Learning Repository. The dataset is the publicly available from the UCI Machine Learning Repository [12]. This dataset initially

contains 569 samples, each representing a breast mass that has been classified as either malignant (M) or benign (B). For each sample thirty numerical features were extracted from each digitized fine-needle aspiration (FNA) image of breast tissue. These numerical features reflect ten key morphological characteristics of the cell nuclei represented in the images. In particular, they include metrics describing radius, texture, perimeter, area, smoothness, compactness, and degree of concavity, number of concave points, symmetry, and fractal dimension. Each of the ten characteristics is presented as three statistics: the mean, the standard error, and the worst value (defined as the largest observed value). Together, these features form a comprehensive morphological profile of each tumor sample, providing a basis for constructing prognostic models aimed at distinguishing benign from malignant cases. No external clinical or imaging data were used. We follow the original feature definitions (mean, standard error, and "worst" of the ten morphology descriptors) and labels. All preprocessing (standardization) was performed within cross-validation folds to prevent information leakage.

Data Preprocessing

At the data preparation stage, we cleaned the data: irrelevant columns, including patient identifiers, and variables with missing values were excluded. During preprocessing stage, only patient identifiers and variables with missing values were excluded. All remaining features were preserved for subsequent feature selection. To form the target variable, categorical diagnosis values were converted to a binary numeric format: malignant tumors ("M") were coded as 1, and benign tumors ("B") were coded as 0. Then, we standardized all numerical features using the z-normalization method using the StandardScaler tool from the scikit-learn library. This allowed us to bring the data to a single scale and avoid the influence of differences in the scale of features on the model training results. This preprocessing step was critical in logistic regression algorithms, which are sensitive to feature magnitude.

Machine Learning Models

Four widely recognized supervised machine learning (ML) classification algorithms, which are logistic regression with L1-regularized penalty (LR) [13] were applied, random forest [14], decision tree [15] and naïve Bayes [16] to determine the most accurate method. These models were selected based on their proven effectiveness in prior breast cancer prediction studies, as well as their general reputation within the machine learning community for handling structured datasets efficiently. Random forest and logistic regression are identified as performing well among models submitted to structured data competitions and applied in real practice in healthcare. For instance, Shwartz-Ziv and Armon reported evidence that logistic regression, random forest, and gradient boosting outperform neural networks in 'tabular data applications' where it is important to remained interpretable, exhibit robustness to overfitting, and are easy to implement [17]. The choice to include logistic regression, tree-based (random forest and decision tree), and naïve Bayes models, different learning paradigms within a single framework and assess which algorithm class is most effective for breast cancer prediction were aimed to be exploreed. In this section, we presented a short description of a mathematical model of these algorithm designs.

L1-regularized logistic regression (LR)

We modeled the probability of malignancy with LR by the L1 norm to promote sparsity and interpretability. Let $\mathbf{x_i} = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ be standardized features for sample i and $y_i \in \{0, 1\}$ the label (malignant=1, benign=0). We model the conditional probability of malignancy by logistic (sigmoid) link:

$$Prob(\boldsymbol{y}_i = 1|\boldsymbol{x}_i) = \boldsymbol{\sigma}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right), \sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

We estimate parameters $\boldsymbol{\beta}$ by minimizing the penalized negative log-likelihood

$$\min_{\beta} \mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n}[y_i \log \widehat{p_i} + (1 - y_i)\log(1 - \widehat{p_i})] + \lambda\|\boldsymbol{\beta}\|_1 \quad (2)$$

with hyperparameter $\lambda = 1/C$ controls sparsity (larger λ-stronger shrinkage) and selected by cross-validation to balance bias-variance and encourage sparse, interpretable solutions. We use an L1-capable solver (*saga*, max_iter=1000). Features are z-scored within CV folds and the L1 penalty mitigates multicollinearity and performs embedded feature selection.

Decision Tree (DT)

A decision tree model is a popular approach for classification and prediction in machine learning. A decision tree partitions the feature space into axis-aligned regions $\{R_l\}_{l=1}^{L}$ and predicts the majority class in each region. Splits are chosen to minimize impurity where impurity is the entropy or Gini (as specified in Table 1):

$$Impurity(x) = Entropy(x) = -\sum_{c\in\{0,1\}} \widehat{p_c}(x) log \widehat{p_c}(x) \text{ Or Gini(t)=1-}\sum_{c\in\{0,1\}} \widehat{p_c}(t)^2 \quad (3)$$

selecting a feature and a threshold that minimizes the weighted post-split impurity. Hyperparameters include maximum depth and minimum samples per leaf.

Random Forest (RF)

Random Forest is a machine learning technique that is used for classification and regression tasks. This model builds many decision trees taking the different subsets of training data. RF is an ensemble of B trees $\{T_b\}_{b=1}^{B}$ trained on bootstrap samples. At each split a random subset of features of size max-features is considered.

$$\widehat{y_i} = mode\left(T_1(\boldsymbol{x_i}), \ldots, T_B(\boldsymbol{x_i})\right),\ \widehat{p_i} = \frac{1}{B}\sum_{b=1}^{B} 1\{T_b(\boldsymbol{x}_i) = 1\} \quad (4)$$

This reduces relative to a single tree. Out-of-bag estimates provide internal validation and variable importance diagnostics.

Naïve Bayes (NB)

Naïve Bayes is a useful machine leaning method for classification tasks. NB assumes conditional independence of features given the class c∈{0, 1}. With Gaussian conditionals for each feature,

$$p(\boldsymbol{x}|\boldsymbol{c}) = \prod_{\boldsymbol{j=1}}^{\boldsymbol{p}} N\left(x_j;\ \mu_{jc}, \sigma_{jc}^2\right),\ \ p(c) = \pi_c, \quad (5)$$

and by Bayes' rule

$$Prob(c|\boldsymbol{x}) = \frac{\pi_c \prod_{\boldsymbol{j=1}}^{\boldsymbol{p}} N\left(x_j; \mu_{jc}, \sigma_{jc}^2\right)}{\sum_{c'\in\{0,1\}} \pi_{c'} \prod_{\boldsymbol{j=1}}^{\boldsymbol{p}} N\left(x_j; \mu_{jc'}, \sigma_{jc'}^2\right)} \quad (6)$$

using per-class means and variances estimated from the training data.

Machine Learning Pipeline

Figure 1 depicts the primary architecture of the ML pipeline developed for this study. We designed the workflow with a systematic, data-driven approach to support thorough evaluation and reliable model selection. A stratified random split was used to first divide the dataset into training (80%) and test (20%) subsets, ensuring that the percentage of benign and malignant cases in each subgroup remained constant. The feature selection, model training, and hyperparameter tuning were performed on the training set via 5-fold cross-validation. Recursive feature elimination (RFE) was used to find the most informative features in this set and enhance model focus. Following feature selection, 5-fold cross-validation was used to hyperparameter tune the logistic regression (with L1 regularization), random forest, and decision tree classifiers. Naïve Bayes was excluded from the hyperparameter tuning step because it does not have adjustable hyperparameters in the optimization techniques. Its performance is primarily

determined by data distribution and the independence assumption, rather than by parameters that can be systematically adjusted through cross-validation. The model performance of each configuration was evaluated using the area under the ROC curve (AUC) as the primary selection criterion metric for selecting the best-performing configuration. Analyses were run in Python (scikit-learn package for models/metrics; an implementation of DeLong's test for AUC comparisons).

We assume that treats samples as independent and identically distributed. To avoid information leakage, every preprocessing step is confined to the cross-validation folds. The Naïve Bayes classifier is used under its standard assumptions of conditional feature independence and approximate normality after standardization. DeLong's test is applied under the regularity conditions required for consistent estimation of the asymptotic variance of AUC differences.

The model that achieved the highest cross-validated AUC on the training set was the one that performed the best. This selected model was then retrained using the full training set with optimal hyperparameters and features. Finally, the held-out test set was used to evaluate the trained model's predicted performance in the real world.

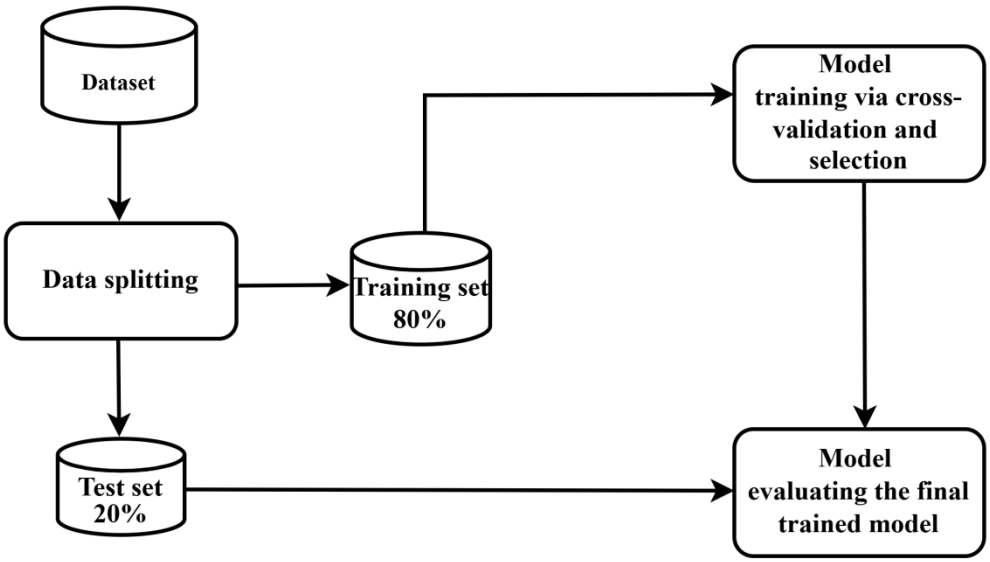

Figure 1. Machine learning pipeline diagram

Model and Feature Selection

A hybrid model and feature selection process to optimize performance and enhance interpretability were used a. This approach allowed us to simultaneously analyze the relationship between individual features and model behavior. The integration of feature selection and model tuning into a single process was aimed at identifying the most informative set of characteristics, as well as the optimal classifier for accurate breast cancer prognosis.

We applied recursive feature elimination (RFE) to feature selection. This embedded feature selection method [18] ranks features according to their importance to the model and progressively removes the least important characteristics. In order to maintain consistency between models and offer a respectable range of predictive power, we selected ten features. For comparability, the same set of 10 features selected by RFE was used to train all classifiers.

After selecting the features, for the model selection we performed hyperparameter tuning for each model using GridSearchCV over the selected features with 5-fold cross-validation (CV). The search space of hyperparameter for each classifier is given in Table 1. We did not tune any

hyperparameters for Naïve Bayes because this model has no significant adjustable parameters in its default form.

Table 1. The hyperparameters space explored via GridSearch as part of model selection using cross-validation

Classification models	**Hyperparameter**	**Hyperparameter Space**
Logistic Regression	Penalty Regularization parameter C	L1 {0.01, 0.1, 1}
Random Forest	Number of estimators Maximum Features Maximum depth	{10, 100, 1000} 'auto', 'sqrt', 'log2' {2, 5, 10, 20, 50}
Decision Tree	Maximum depth Criterion Minimum samples per leaf	{1, 2, 10} 'gini', 'entropy' {1, 2, 10}
Naive Bayes	-	-

Only the features chosen through the RFE procedure were used to train each model. Then, using the highest AUC score obtained after 5-fold cross-validation, we determined which configuration of each algorithm performed the best. By using this method, we were able to compare classifiers while maintaining feature relevance in the final prediction models.

Model Evaluation

The best model performance on training set was evaluated based on the held-out test set (20%). This final step was essential for understanding how well the models generalized beyond the training environment. We used the following evaluation metrics in terms of the AUC, accuracy, sensitivity and specificity.

Results

Performance of ML models: We used 5-fold cross-validation to estimate how well each model could separate benign from malignant cases, and those AUC scores are shown in Table 2. These values helped us get a clearer idea of each model's behavior during training and were one of the key things we looked at when comparing their overall performance.

Table 2. AUC results from 5-fold cross-validation for each classifier, across different hyperparameter settings

Models	**AUC**	**Accuracy**	**Selected Hyperparameters**
Logistic Regression	**0.994**	**0.964**	{'C': 1, 'penalty': 'L1'}
Random Forest	0.987	0.962	{'max_depth': 5, 'n_estimators': 10', 'max_features': 'sqrt'}
Decision Tree	0.942	0.938	{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2}
Naive Bayes	0.985	0.938	-

The results in Table 2 shows that the logistic regression classifier achieving an AUC of 0.994 and an accuracy of 0.964, which reflects strong predictive capability. The random forest classifier follows with a slightly lower AUC of 0.987. Naïve Bayes and the decision tree classifiers achieved AUC scores of 0.985 and 0.942, respectively. Although up to 1000 estimators were tested, the best results were achieved with 10 trees. We found out that adding more trees did not improve performance. This indicates excellent ability to discriminate between

benign and malignant cases at different classification thresholds. These results highlight the effectiveness of linear and ensemble-based methods in medical datasets.

Feature statistics for benign and malignant classes: Table 3 lists the average values for the ten features we picked out in our analysis, shown separately for benign and malignant tumors. The numbers make it pretty clear that the two groups differ quite a bit. In most cases, malignant tumors tend to have higher average values than benign ones, which show a clear and pretty consistent difference between the two groups. For example, features such as *concave points_mean*, *radius_worst*, *perimeter_worst*, and *concave points_worst* stand out the gaps in their values between benign and malignant cases are particularly large difference. These kinds of shifts help the model figure out which features really matter during classification. In general, these results show that the features are highly informative and contribute model's ability to make reliable distinctions between benign and malignant tumors.

Table 3. Mean values of selected features in benign and malignant tumor groups

Feature	**Benign Mean**	**Mlignant Mean**
concave points_mean	-0.598465	1.007793
radius_se	-0.437038	0.735956
area_se	-0.422475	0.711432
compactness_se	-0.225788	0.380218
radius_worst	-0.598342	1.007585
texture_worst	-0.352093	0.592912
perimeter_worst	-0.603320	1.015969
area_worst	-0.565492	0.952267
concavity_worst	-0.508301	0.855960
concave points_worst	-0.611529	1.029791

Statistical Analysis: Differences between AUCs were evaluated using DeLong's test with Holm correction for correlated ROC curves. It helps us to compare correlated ROC AUCs on the same test set. Pairwise statistical comparisons using DeLong's test with Holm correction for multiple testing indicated that the differences between logistic regression and random forest (p=0.21) or naïve Bayes (p=0.19) were not statistically significant (adjusted p-value>0.05). In contrast, the decision tree demonstrated a significantly lower AUC compared with logistic regression (adjusted p=0.04, *p-value*<0.05). These findings confirm the suitability of models such as logistic regression, which achieved state-of-the-art performance. Table 4 summarizes the area under the ROC curve (AUC) with 95% confidence intervals for each classifier, together with the results of pairwise statistical comparisons against logistic regression using DeLong's test with Holm correction.

Table 4. ROC AUC (95% CI) by model and DeLong test *p-values* vs logistic regression (Holm-adjusted)

Model	**AUC (95% CI)**	*p-value* vs LR (Holm-adj)
Logistic Regression	0.996 (0.983-0.999)	-
Random Forest	0.987 (0.969–0.998)	0.21
Decision Tree	0.942 (0.899–0.974)	0.04*
Naïve Bayes	0.985 (0.966–0.996)	0.19

(* - statistically significant difference after Holm's correction)

Best performing model: We selected the best logistic regression (LR) model based on training results and then tested it on a held-out dataset to see how well it would perform on new

cases. According to Table 5, the LR model handled results quite well across all evaluation metrics. The model reached an AUC of 0.996 and a high accuracy of 0.973. Also, it demonstrates excellent sensitivity and high specificity. It is reflecting its effectiveness in correctly classifying benign cases.

Table 5. Best performing LR model on the test set

Models	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
Logistic Regression	0.996	0.973	0.952	0.986

Receiver Operating Characteristic (ROC) curves was plotted for all the models to get a clearer sense of how well they separate the two classes. Figure presents the ROC curves used to compare model performance based on 5-fold cross-validation applied on the training set.

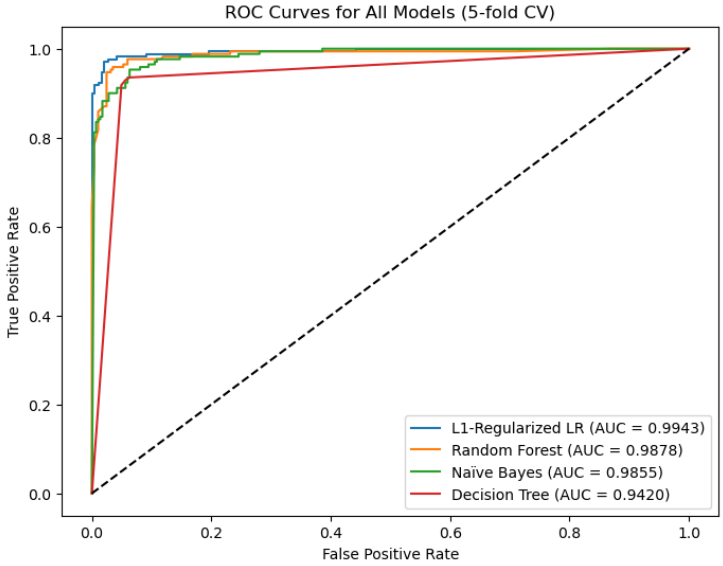

Figure 2. ROC curves comparing the performance of each classifier based on 5-fold cross-validation on the training data

The ROC curves revealed strong discriminative performance, with Logistic Regression (L1-Regularized) achieving the highest area under the curve (AUC=0.9943), Random Forest (AUC=0.9878), Naïve Bayes (AUC=0.9855), and Decision Tree (AUC=0.9420). The ROC curves highlight the superior predictive ability of the L1-regularized Logistic Regression model compared to the other classifiers.

Figure 3 represents a confusion matrix that we plotted only for best best-performing logistic regression model to understand the insight into the prediction outcomes. We can interpret not only the model's overall accuracy but also the frequency of correct and incorrect predictions to understand its practical diagnostic potential.

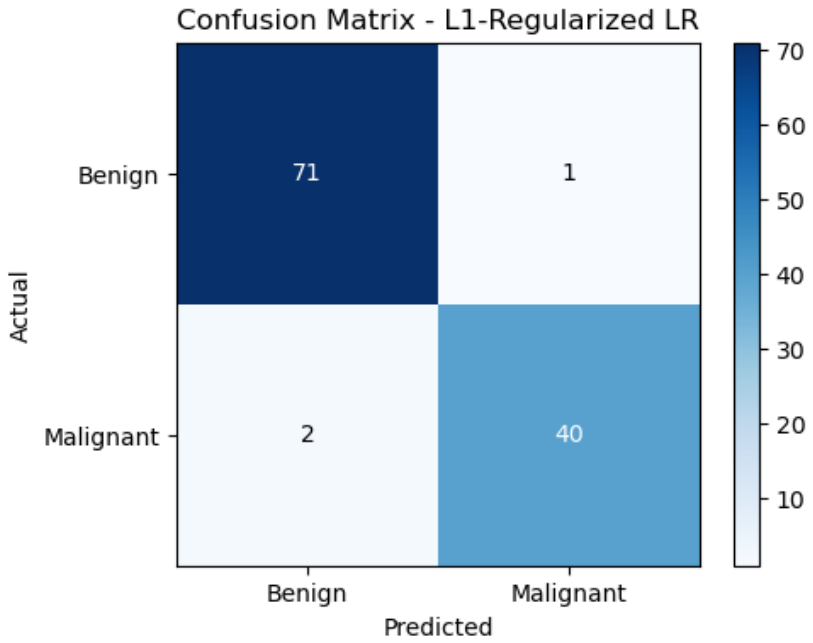

Figure 3. Confusion matrix for the L1-requlararized logistic regression model evaluated on the independent test set

The classification results of the confusion matrix confirmed its robust performance, where true positives (TP) show 40 malignant tumors, true negatives (TN) - 71 of benign tumors, false positives (FP) - 1 of benign tumors, and false negatives (FN) - 2 of malignant tumors. High sensitivity and specificity indicate that the model can be trusted to distinguish benign from malignant breast cancer cases in clinical applications.

Discussion

In this study, machine learning techniques show real promise in detecting breast cancer earlier and more accurately. We focused on interpretable models, particularly L1-regularized logistic regression, which was valuable and critical to predict breast cancer. Among the four classifiers we evaluated, logistic regression with L1 regularization turned out to perform best overall. It consistently showed both high sensitivity and specificity, with AUC reaching 99.4% on the training data and 99.6% on the test set. These outcomes generally confirm our expectation that, when properly tuned and regularized, logistic regression can provide strong and generalizable diagnostic performance.

Our results are not only consistent with our expectations but also consistent with previous studies using the same Breast Cancer Wisconsin (Diagnostic) dataset. For example, Agarap [19] tested models such as multilayer perceptron (MLP) [20] and support vector machines (SVM) [21], MLP algorithm is reaching test accuracies of $\approx$ 99.04 % with a 70/30 train-test split. Similarly, Entezari [22] discovered that SVM achieved 98% of accuracy and AUC of 99%, outperforming other models such as k-nearest neighbours (KNN) [23] and Bayesian logistic regression [24]. Using higher-order probabilistic perceptrons, Cowsik and Clark [25] reported an accuracy of ~97%. Because the model assumes feature independence, which isn't always realistic in medical datasets, Naïve Bayes performed marginally worse in their experiments, at about 95%. Ghosh et al. [26] employed XGBoost [27] in their more recent research and reported an accuracy of almost 97.7% on the same dataset. Murty et al. [28] used a hybrid deep learning framework and combined WDBC and CBIS-DDSM. The CNN model produced strong AUC values, reportedly above 0.96. In a similar vein, Aamir and Rahim [29] assessed several classifiers, including SVM, random forest, gradient boosting, and a hybrid MLP. Their hybrid model, which used a connection-based feature selection method and 5-fold cross-validation, performed the best, achieving 99.12% accuracy.

SVM and a feature selection technique were combined in an earlier but noteworthy study by Akay [30], which demonstrated that up to 99.51% accuracy, could be attained using only five top features. This demonstrates how feature reduction can result in more effective models and better performance.

While these studies highlight the power of ensemble methods. Our logistic regression model with L1 regularization was performed at the level of complex models but was easier to interpret. More interpretable models are also supported in the wider literature. For example, Shwartz-Ziv and Armon [17] showed that deep learning models often do not outperform linear ones when working with tabular data. Similarly, Gupta et al. [2] demonstrated that logistic regression and decision trees suggest competitive accuracy with the additional benefit of interpretability. Our model's ability to identify and focus on the most relevant features, facilitated by RFE, further reduced noise and overfitting, improving robustness and generalization.s, facilitated by RFE, further reduced noise and overfitting, enhancing robustness and generalization.

Furthermore, we observed strong performance from the Random Forest and Naïve Bayes classifiers, although they lacked the diagnostic accuracy demonstrated by Logistic Regression. These observations are consistent with the conclusions drawn by Lu et al. [10], who found that Random Forest performed well when different data types were combined. However, they also pointed out that for datasets based purely on clinical variables, well-tuned linear models can be just as effective.

In our case, using L1 regularization made the model stay sparse, so it became clearer which variables were affecting the predictions. However, in clinical practice, interpretability is important to ensure responsible use of AI.

We applied an RFE approach to improve the performance of the model on relevant variables. This RFE method was proposed by Guyon et al. [18] to select genes. We were able to choose the top ten features in our case due to RFE. This reduced the number of less useful inputs and helped the model work better on unseen data. This not only reduced irrelevant input but also helped the model generalize better. Similar strategies have been used before in studies focused on cancer risk stratification [16]. To ensure that observed differences in performance were meaningful, we performed pairwise statistical comparisons of AUCs using DeLong’s test with Holm correction. These analyses confirmed that logistic regression performed significantly better than the decision tree, while differences between logistic regression and random forest or naïve Bayes were not statistically significant.

The main limitations of this study are that the dataset used is relatively small and mostly includes patients from a similar demographic background. We do not know how the model would perform on more diverse or larger populations. These results obtained on a relatively small and homogeneous dataset. Performance may be lower when applied to larger, more diverse clinical populations, and further external validation is required. Our analysis only focused on structured variables. Using different types of data, such as imaging, genomic information, or even electronic health records, may help improve both the accuracy and the usefulness of the model.

In the future, it would be important to test the model on data from different research clinics to assess whether it holds up in different settings. We also see potential in using ensemble methods or combining different models. It could make predictions more stable. Also, using explainability AI techniques could help clinicians better understand how the model makes its decisions and if such tools will be used in real practice.

Conclusion

In analysing breast cancer diagnostic data, various machine learning approaches can be applied. A key challenge is identifying accurate, interpretable, and clinically reliable models. In this study, we evaluated four classifiers as L1-regularized logistic regression, random forest,

decision tree, and naïve bayes on the Breast Cancer Wisconsin (Diagnostic) dataset. We compared model performance in terms of AUC, accuracy, sensitivity, and specificity using recursive feature elimination and 5-fold cross-validation.

Our results showed that the L1-regularized logistic regression did surprisingly well with an AUC that was close to perfect, and it even outperformed some of the more advanced models. We also found that using feature selection helped models built on selected features work better than those using everything. However, this study focused only on a small dataset of classifiers. It should try other types of models, maybe add imaging or genomic data, and also test on a larger dataset to see if the findings hold up.

In conclusion, this work shows that when combined with explainable machine learning, information technology has real potential to help build smarter diagnostic tools for use in healthcare.

Acknowledgment

This research was supported by grant AP19679717 from the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan awarded to Z.R., Y.B., A.M., and A.U. The funding body had no involvement in the study's design, data analysis, decision to publish, or manuscript preparation. B.M. received support from the Ministry of Science and Higher Education of the Republic of Kazakhstan through grants BR24992841 and BR24993023. All authors had unrestricted access to the study data, with the lead authors (Z.R., Y.B., B.M., A.U.) holding final authority over the decision to submit the manuscript for publication.

References

[1] Siegel, R. L., Miller, K. D., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians, 73*(1), 17-48. https://doi.org/10.3322/caac.21763

[2] Gupta, M., Jain, R., Solanki, A., & Al-Turjman, F. (Eds.). (2021). Cancer Prediction for Industrial IoT 4.0: *A Machine Learning Perspective (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003185604

[3] Duffy, S. W., Vulkan, D., Cuckle, H., Parmar, D., Sheikh, S., Smith, R. A., & Evans, A. (2020). Effect of mammographic screening from age 40 years on breast cancer mortality. *The Lancet Oncology*, *21*(1), 113-122. https://doi.org/10.1016/S1470-2045(19)30721-5

[4] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

[5] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & van der Laak, J. A. W. M. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

[6] Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Radiology*, *294*(2), 350-367. https://doi.org/10.1002/jmri.26534

[7] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

[8] Zhang, T., Tan, T., Han, L., et al. (2023). Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. *npj Breast Cancer, 9*, Article 16. https://doi.org/10.1038/s41523-023-00517-2

[9] Mu, J., Nazar, A., Ali, M. A., & Hussain, A. (2025). Integrating machine learning with OMICs data for early detection in breast cancer. *Gene Reports, 41*, 102325. https://doi.org/10.1016/j.genrep.2025.102325

[10] Lu, C., Wang, J., Zhang, H., & Wang, S. (2022). Integrating histopathological images and genomic data for breast cancer subtype classification. *Frontiers in Oncology, 12*, 928763. https://doi.org/10.3389/fonc.2022.928763
[11] Hussain, S., Lafarga-Osuna, Y., Ali, M., Naseem, U., Ahmed, M., & Tamez-Peña, J. G. (2023). Deep learning, radiomics and radiogenomics applications in the digital breast tomosynthesis: A systematic review. *BMC Bioinformatics, 24*, Article 401. https://doi.org/10.1186/s12859-023-05515-6
[12] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast Cancer Wisconsin (Diagnostic)* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B
[13] Demir-Kavuk, O., Kamada, M., Akutsu, T., & Knapp, E. W. (2011). Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinformatics, 12*, Article 412. https://doi.org/10.1186/1471-2105-12-412
[14] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[15] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. *Chapman and Hall/CRC*. https://doi.org/10.1201/9781315139470
[16] Webb, G. I. (2011). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning*. Springer. https://doi.org/10.1007/978-0-387-30164-8_576
[17] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
[18] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797
[19] Agarap, A. F. (2017). On breast cancer detection: An application of machine learning algorithms on the Wisconsin Diagnostic Dataset [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1711.07831
[20] Bourlard, H. A., & Morgan, N. (1994). Multilayer perceptrons. In *Connectionist speech recognition: A hybrid approach* (The Springer International Series in Engineering and Computer Science, Vol. 247, pp. 59–80). Springer. https://doi.org/10.1007/978-1-4615-3210-1_4
[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
[22] Entezari, R. (2018). Breast cancer diagnosis via classification algorithms [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1807.01334
[23] Fix, E., & Hodges, J. L., Jr. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique, 57*(3), 238–247. https://doi.org/10.2307/1403797
[24] Gosho, M., Ohigashi, T., Nagashima, K., Ito, Y., & Maruo, K. (2023). Bias in odds ratios from logistic regression methods with sparse data sets. *Journal of Epidemiology, 33*(6), 265–275. https://doi.org/10.2188/jea.JE20210089
[25] Cowsik, A., & Clark, J. W. (2019). Breast cancer diagnosis by higher-order probabilistic perceptrons [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1912.06969
[26] Ghosh, P. (2022). Breast Cancer Wisconsin (Diagnostic) prediction. *International Journal of Science and Research (IJSR), 11*(5), 178–185. https://doi.org/10.21275/SR22501213650
[27] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
[28] Murty, P. S. R. C., Anuradha, C., Naidu, P. A., et al. (2024). Integrative hybrid deep learning for enhanced breast cancer diagnosis: Leveraging the Wisconsin Breast Cancer Database and the CBIS-DDSM dataset. *Scientific Reports, 14*, Article 26287. https://doi.org/10.1038/s41598-024-74305-8

[29] Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., Khan, M. A., Khan, K., & Ahmad, J. (2022). Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine, 2022*, Article 5869529. https://doi.org/10.1155/2022/5869529

[30] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*(2), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009