Scientific Journal of Informatics
Vol. 13, No. 1, Feb 2026

P-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Optimizing Early Breast Cancer Classification Using Hybrid SVM-ANN with Ridge Embedded Feature Selection

Sigit Priyanta[1*], Dita Ria Selvyana[2], Aulia Salsabila[3]
[1,3]Department of Computer Science and Electronics, Faculty of Mathematics and Computer Sciences, Universitas Gadjah Mada
[2]Department of Internal Medicine, Faculty of Medicine and Health Sciences, Universitas Muhammadiyah Yogyakarta

Abstract.
Purpose: This study aims to enhance early breast cancer detection by systematically evaluating multiple machine learning (ML) algorithms and feature selection strategies. The goal is to identify the most effective combination of classifiers and feature selection methods for accurately distinguishing malignant from benign breast tumors, thereby improving diagnostic reliability and clinical decision support.
Method: The Wisconsin Breast Cancer Dataset containing 699 samples described by nine diagnostic features was used. Tumor classes were encoded as 0 (malignant) and 1 (benign). The analysis was conducted in two stages. First, five ML algorithms—K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN), and a hybrid SVM–ANN—were evaluated to establish baseline performance. Second, two feature selection approaches (wrapper and embedded) were applied to four ML models and the optimized hybrid classifier. The embedded approach employed Ridge-based feature selection to identify the most discriminative attributes and improve model generalization.
Results: The hybrid SVM–ANN combined with Ridge Embedded feature selection achieved the best performance, with an accuracy of 97.86%, precision of 96.5%, recall of 96.5%, and an F1-score of 96%. This configuration outperformed all other algorithms and feature selection techniques, affirming the effectiveness of hybrid integration and embedded feature optimization.
Novelty: The novelty lies in the integration of an SVM–ANN hybrid model with Ridge-based embedded feature selection for breast cancer classification. Unlike prior works that rely primarily on conventional filter or wrapper techniques, this approach demonstrates superior accuracy and robustness. The proposed framework provides a promising pathway for developing more reliable ML-based diagnostic tools in oncology.

Keywords: Breast cancer, Machine learning, Feature selection, SVM-ANN optimization.
Received May 2020 / **Revised** November 2020 / **Accepted** March 2021

INTRODUCTION

Breast cancer represents an abnormal and harmful proliferation of cells arising within the tissues of the mammary glands, primarily originating from either the lobules or the ducts. Global estimates suggest that cancer incidence may escalate from about 14 million to roughly 22 million within the coming twenty years, and this number is projected to continue growing each year progressively [1], [2], [3]. This upward trend highlights the increasing the critical importance of detecting the disease at an early stage alongside and effective clinical management.

The population level impact associated with breast cancer is also reflected in recent global and national statistics. According to the Global Cancer Observatory (GCO) in 2020, breast cancer ranks first as the largest contributor to cases of 35 types of cancer. There are around 65,858 cases, or 16.6% of the total 396,914 cancer cases in Indonesia. Breast cancer also contributes to 22,430 cases of the total 234,511 cases of cancer deaths and is the second leading cause of death after lung cancer in Indonesia [4].

Based on these problems, early detection of breast cancer is essential as an initial step in treating breast cancer [5]. Malignant and Benign are two types of tumors that are widely used in determining breast cancer. In the medical procedure, patients undergo a series of tests, namely ultrasound, biopsy, and mammography, based on the variation of breast cancer symptoms experienced. One widely used method is biopsy, by extracting a tissue or cell sample from the body for further analysis in the laboratory. The breast tissue or cell samples collected are obtained using the Fine Needle Aspirate (FNA) procedure. FNA is applied to the part suspected of having malignant tumor growth.

*Corresponding author.
Email addresses: seagatejogja@ugm.ac.id (Priyanta)
DOI: [10.15294/sji.v13i1.36676](https://doi.org/10.15294/sji.v13i1.36676)

Microscopic examination yields nine quantitative descriptors, which encompass measurements such as clump thickness, variations in cell size and shape, levels of marginal adhesion, the size of individual epithelial cells, presence of bare nuclei, chromatin texture, nucleolar features, and the rate of mitosis [6]

Features obtained from Fine Needle Aspiration (FNA) are essential indicators used to assess the probability that a given sample corresponds to malignant tumor growth. Malignant tumors generally spread more aggressively than benign tumors, making early identification crucial [7]. Therefore, automatically detecting and distinguishing benign from malignant tumors is a critical first step to ensure timely treatment and reduce breast cancer–related mortality.

Automated detection of breast cancer can be effectively performed through classification techniques [5]. Machine learning (ML) and data mining (DM) methods are widely used by researchers to support this because they can process and interpret intricate medical information with high effectiveness [8]. ML-based systems offer several advantages, including reducing manual workload and automatically categorizing clinical images and textual data, making them a promising tool for future cancer screening applications [9] [10].

Various machine learning (ML) techniques have been extensively utilized to forecast breast cancer outcomes, including models such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Networks (ANNs) [5], [11]. [5] evaluated multiple ML algorithms applied to the Wisconsin Breast Cancer (WBC) dataset—among them Random Forest (RF), Decision Trees, Naïve Bayes, LR, SVM, and KNN—and reported that SVM and RF achieved the highest performance, achieving an accuracy rate of 96.5%. Similarly, Singhal et al. [7] analyzed five algorithms (SVM, RF, LR, C4.5, and KNN) using the same dataset and found that SVM produced the best accuracy of 97.19%, although their study did not include precision, recall, or F1-score metrics. When examining lung cancer detection conducted with the WBC dataset [12], compared SVM and ANN and concluded that SVM consistently delivered superior performance, with average accuracy, precision, recall, and ROC scores around 96%.

Findings from these studies indicate that SVM is the most frequently used ML method and consistently achieves top performance in breast cancer classification. Further improvements have been demonstrated by [13], who integrated the C4.5 algorithm with KNN imputation and Relief-based the selection of relevant features. Their hybrid approach achieved an accuracy of 98.57%, outperforming established benchmarks such as PSO-C4.5, EBL-RBFNN, Gaussian Naïve Bayes, and t-SNE–based models. This result underscores how integrating feature-selection strategies with hybrid ML methods can significantly improve diagnostic precision.

Support Vector Machines operate by identifying a decision boundary that maximally separates samples belonging to different classes. [7] [14]. The SVM algorithm can work with many features by obtaining hyperplanes in multidimensional space [15]. SVM can be used as an optimization algorithm, namely, reducing the dimensions of the dataset [16]. ANNs are computational models inspired by the structural and functional characteristics of the human brain. Groups of interconnected neurons are arranged into layers—typically an input layer, one or more hidden layers, and an output layer, which together form deep machine learning, also known as Deep Learning (DL) [17]. SVM Optimized is a Hybrid ML algorithm that utilizes SVM for feature extraction to reduce dimensions and ANN for classifying breast cancer, thereby improving prediction results and achieving high model accuracy.

Numerous prior investigations have applied hybrid machine learning approaches to classify breast cancer cases using the Wisconsin Breast Cancer dataset, demonstrating the effectiveness of combining dimensionality reduction with advanced classifiers. Study applied SVM for feature extraction to reduce dataset dimensionality, followed by a CNN classifier, achieving average accuracy, specificity, and sensitivity of 95%. [15] similarly used SVM for dimensionality reduction and ANN for classification, reporting strong performance with an accuracy of 94.5%, as well as precision, recall, and F1-score metrics. [18] employed Principal Component Analysis (PCA) to reduce dimensions before applying ANN, attaining average accuracy, sensitivity, and specificity of 96%. Although these studies produced promising results in breast cancer prediction, they relied exclusively on dimensionality reduction and did not incorporate feature selection during preprocessing. While dimensionality reduction and feature selection both aim to reduce the number of attributes in a dataset, they differ fundamentally: dimensionality reduction generates new transformed feature combinations, whereas feature selection retains the original attributes and selects only the most relevant subset without modification [18].

Feature selection is a process of identifying the subset of attributes most critical to the task taken from the training data to enhance how efficiently and accurately the model performs classification. Its primary aim is to decrease the overall number of dimensions represented in the dataset—such as vocabulary size in text classification—thereby making the learning process more efficient [19]. According to [19], selecting informative features is a critical component of machine learning, as classification algorithms generally consider all available features even though not all of them contribute meaningfully to the predictive task. In the literature, the procedure for determining which attributes contribute the most meaningful information is often organized into three methodological classes: filters, wrappers, and approaches that incorporate feature selection directly into the learning model [20]. The wrapper method selects subsets of attributes and evaluates them using a classifier such as Decision Tree, Naïve Bayes, or SVM, while the embedded method performs feature selection during model construction as part of the learning algorithm itself, eliminating the need for repeated execution [20]. Applying feature selection in breast cancer classification allows the model to focus on the most informative attributes, leading to improved accuracy and overall predictive performance.

Several previous studies that applied feature selection in breast cancer classification are [21], which compared feature selection methods using embedded and chi-squared methods on the Wisconsin Breast Cancer (WBC) dataset, achieving the highest accuracy with the embedded method at 96% and the chi-squared method at 94%. However, this study did not optimize the ML algorithm for classification. [22] conducted feature selection using the wrapper method with the SVM algorithm on the WBC dataset, and classification using the SVM algorithm yielded results with 95% accuracy. In contrast, classification results using the SVM algorithm without feature selection achieved a 96% accuracy. However, this study did not optimize the SVM. [23]applied feature selection using the wrapper method on the WBC dataset, namely LR, LSVM, and QSVM, and obtained the highest accuracy results using QSVM of 96% but did not optimize the SVM algorithm.

In this study, an analysis will be carried out with the Wisconsin Breast Cancer (WBC) dataset as the primary source of data, which comprises 9 features and serves as a reference dataset for breast cancer detection and prediction. This research is divided into two stages. The the initial phase involves evaluating five machine-learning algorithms, namely KNN, LR, SVM, ANN, and the hybrid ML algorithm, namely SVM-ANN, optimized to determine their capability in distinguishing benign from malignant breast tumors. The the subsequent phase of this research is to apply feature selection using two methods, namely the wrapper and embedded methods, using 4 ML algorithms, and the optimized ML SVM-ANN algorithm, and compare the performance results of each method and determine which method provides the most optimal performance results in classifying breast cancer. The performance metrics applied to assess classification success include accuracy, precision, recall, the F1-score, and the ROC curve.

METHODS

This section outlines the process stages undertaken in this research, including data preparation, the selection of relevant features, classification, and model performance evaluation. The flowchart of the breast cancer prediction process in this study is presented in Figure 1.

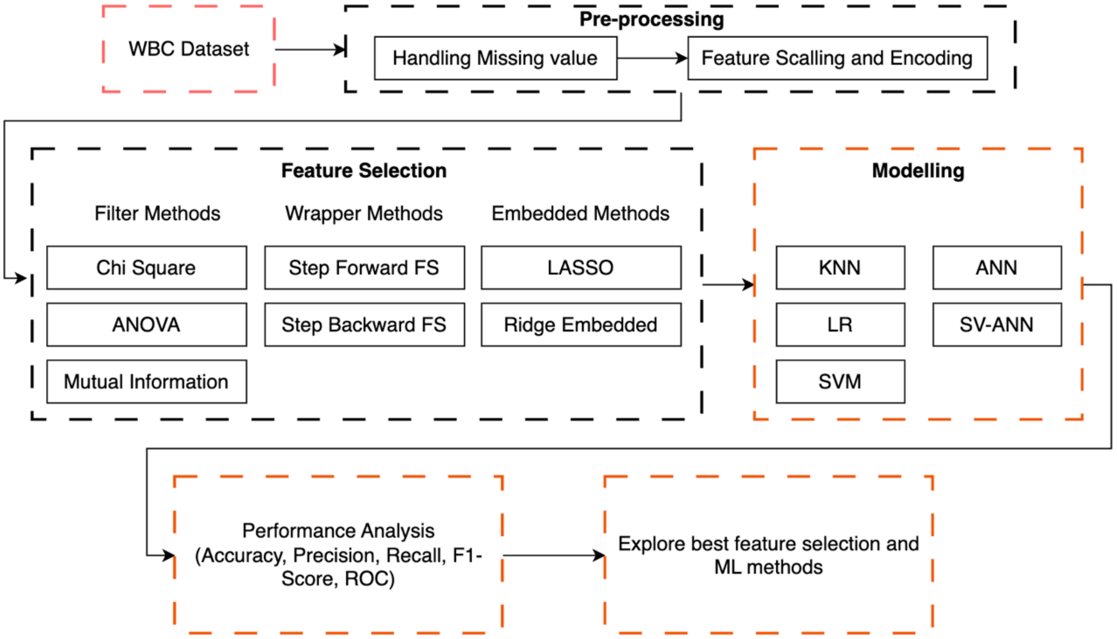


Figure 1. Cancer prediction model development flow

Dataset Description

This research employs the Wisconsin Breast Cancer dataset as the primary source of breast cancer data, which comprises 699 instances with 9 features. The dataset, with 9 columns, is divided into two groups: the first group contains patient data, while the second group contains diagnostic findings. Malignant and benign values are categorized as 0 and 1, respectively, in this study. Patients were categorized as having benign tumors in the dataset, comprising 458 objects, and patients were categorized as having malignant tumors in the dataset, comprising 241 objects. A summary of the dataset's feature attributes is presented in Table 1.

Table 1. WBC dataset features description

WBC Features	Range of values
Sample code ID number	Special ID Number
Clump thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal adhesion	1-10
Single epithelial cell size	1-10
Bare nuclei	1-10 (missing values "?")
Bland chromatin	1-10
Normal nucleoli	1-10
Mitoses	1-10
Class	2 'Benign' and 4 'Malignant'

Preprocessing

1. Handling Missing Data

The dataset includes several attributes with missing entries, for which class-specific mean values are computed for every feature. Furthermore, and these computed class-level means are used to fill the missing entries accordingly based on the appropriate class value. This method will be used and the results will be evaluated [22].

2. Feature Scaling and Categorization

After the missing entries have been addressed, the robust scaler is then employed to normalize the numerical attributes. The choice of Robust Scaler is based on its ability to handle outliers offering improved resilience to outliers compared with standard approaches like Min–Max scaling or Z-score normalization [24]. It operates by subtracting the median of the feature and then scaling it by the interquartile range (IQR). The equation used in Robust Scaler can be seen in Equation (1) [24] below.

$$x_{scaled} = \frac{x - Q2(x)}{Q3(x) - Q1(x)} \quad (1)$$

Where x_scaled is the value of feature x resulting from normalization using Robust Scaler, $Q2(x)$ is the median or middle value of feature x, and $Q3(x)$ is the third quartile of feature x [24].

Feature Selection

Refers to selecting a subset of variables that provide the most meaningful information from the full dataset. Its main purpose is to boost model performance by decreasing the dimensionality of the input features [25]. Feature selection has consistently and has consistently demonstrated benefits in increasing both the accuracy and robustness of cancer-classification systems. [26] Demonstrated that Backward Elimination enhanced the performance of Logistic Regression and SVM. Similarly, Purwinarko et al. [13]employed Relief feature selection in conjunction with C4.5 and KNN imputation, highlighting the effectiveness of targeted feature reduction. Pramesti et al. [27], through the NASNet Mobile CNN model, further emphasized the value of strong feature extraction, even in image-based classification. [28] also showed that the performance limitations of Gradient Boosting Machines were largely influenced by imbalanced and uninformative features, reinforcing the need for more robust feature selection. Building on these findings, the present study employs Ridge Embedded Feature Selection within a Hybrid SVM–ANN framework, enabling the extraction of the most relevant features while addressing multicollinearity and noise. This integration strengthens model stability and boosts classification accuracy for early breast cancer detection.]]] This study explores three different feature selection (FS) techniques, namely filter, wrapper, and embedded, to select the optimal number of features.

1. Filter Method:

a. Chi-Square (CS)

To identify the most relevant attributesthe Chi-square statistic evaluates how strongly each feature is associated with the target output. It is based on the Chi-squared statistic, which tests whether the pattern of observed values is consistent with what would be expected under the hypothesized model. The Chi-Square test examines how much observed frequencies differ from their expected counterparts [29]. The relationship is calculated using the formula (2).

$$\mathrm{x}_{\mathrm{c}}^{2} = \frac{\sum(\mathrm{O_i} - \mathrm{E_i})^2}{\mathrm{E_i}} \quad (2)$$

Where c is where d represents the degrees of freedom, O denotes the observed counts, and E signifies the expected counts [29].

b. Analysis of Variance (ANOVA)

ANOVA is employed to determine whether the means of multiple groups differ beyond what random variation would suggest. It relies on the F-test to evaluate whether variation between groups exceeds variation within groups. If there is no difference between two groups whose values are the same, then the result of the ANOVA F-ratio will be close to 1 [30].

c. Mutual Information (MI)

Mutual information quantifies how much knowledge of one variable reduces uncertainty about another. This value is MI becomes zero only when the variables show complete independence, while larger values reflect stronger dependence [31].

2. Wrapper Method

The wrapper feature selection method outperforms other existing methods, such as the filter method. This method finds the most “useful” features and performs optimal feature selection for the learning algorithm [32]. This study employs two wrapper methods, namely FFS and BFE.

a. Forward Feature Selection (FFS)

Forward Feature Selection (FFS) incrementally constructs a model by introducing one feature at a time, choosing those that most improve predictive performance. This method the process begins with no selected features, gradually adding the most informative ones until the predefined stopping rule is satisfied. This method is particularly useful when dealing with data that has many features, as it incrementally builds a model based on the most informative features. This process involves assessing new features, evaluating feature combinations, and selecting the optimal subset of features that contribute most to model accuracy [33].

b. Backward Feature Selection (BFE)

Backward Elimination removes features sequentially, eliminating them step-by-step from the model, assessing how each removal affects performance and continuing the process until no additional gains occur, ultimately isolating the most predictive subset. [33].

3. Embedded Method:

a. Lasso

Lasso regression belongs to the family of linear models that incorporate regularization techniques. This regression It introduces a penalty component into the loss function, relying on L1 regularization rather than L2, which is why it is commonly referred to as the L1 norm or L1 regularization. This component represents representing the sum of absolute coefficient values multiplied by the regularization parameter λ (lambda) [34]. In general, lasso regression is considered a shrinkage method; some features can be as small as zero, which results in their removal. Therefore, this shrinkage method can be applied in feature selection [35].

b. Ridge Embedded Method

Ridge regression represents a regularized variant of the conventional linear regression model. This means it is a variation of the standard linear regression model that includes a regularized term in the cost function. The goal is to prevent overfitting. Ridge regression adds an L2 regularization term to the linear equation. That is why it is also known as L2 Regularization or L2 Norm. Ridge regression is a shrinkage method that can be applied in feature selection [35].

Modelling

1. Hybrid SVM–ANN Model

The Hybrid SVM–ANN Model is an integrated machine learning approach that combines the strengths of SVMs and ANNs within a unified framework to enhance classification performance. The dataset is first processed using the SVM classifier, which provides strong baseline predictive performance. The SVM-generated predictions are subsequently passed into an ANN to perform the final breast cancer prediction. Therefore, Integrating SVM with ANN enhances predictive accuracy by leveraging both models' strengths [15]. The sequence of operations in the SVM, ANN method are illustrated in Figure 2 as follows.

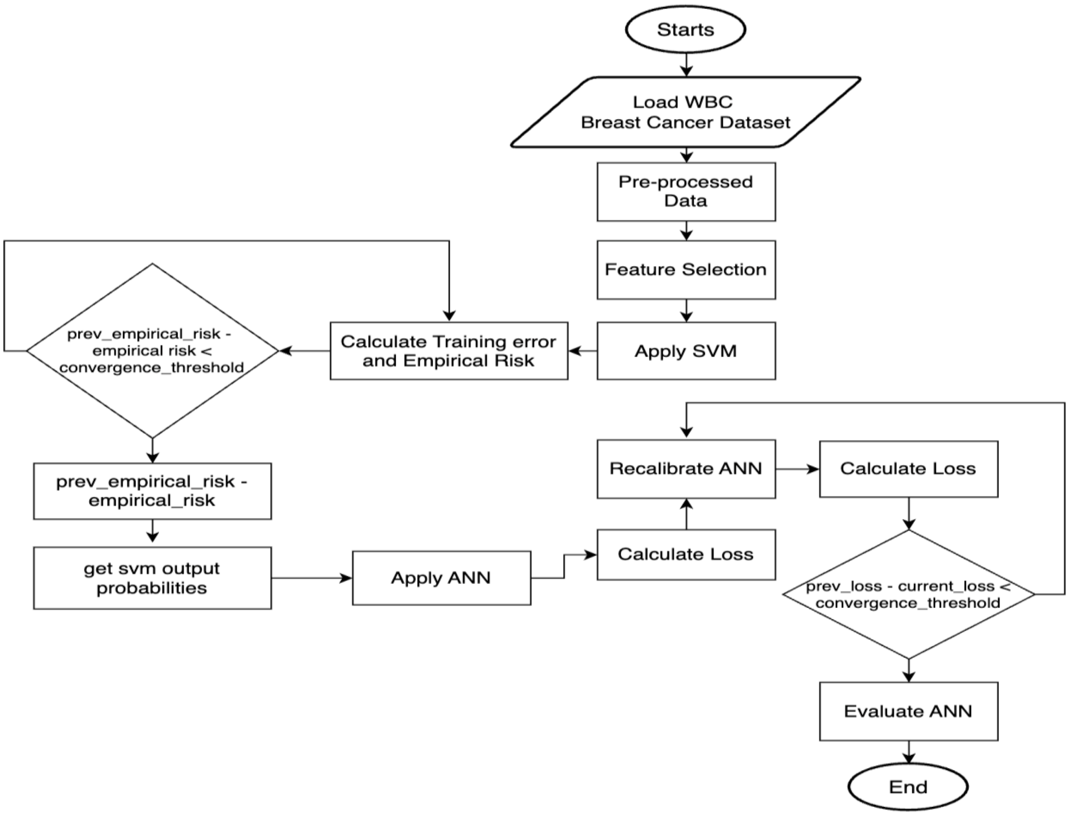

Starts
Load WBC Breast Cancer Dataset
Pre-processed Data
Feature Selection
Apply SVM
Calculate Training error and Empirical Risk
prev_empirical_risk - empirical risk < convergence_threshold
prev_empirical_risk - empirical_risk
get svm output probabilities
Apply ANN
Calculate Loss
Recalibrate ANN
Calculate Loss
prev_loss - current_loss < convergence_threshold
Evaluate ANN
End

Figure 2. SVM-ANN breast cancer detection development flow

In Figure 2, the flow of the optimized SVM-ANN algorithm is illustrated. After passing through the stages of each feature selection method, the features selected by each method are classified using SVM, followed by the calculation of the error value and empirical risk value. Empirical risk is a concept from statistical learning theory that measures how well a machine learning model behaves across its training set. Empirical Risk represents the average error of all training data and measures the error or "risk" of the model [36]. Iterations are carried out if the empirical risk value has reached the convergence limit, then end the training and get the SVM probability output, then the probability output is used as input to the ANN as training data and labels, then the loss or error and recalibrate the ANN layer to adjust the model so that it can be adaptive, calculate the loss and the results obtained are evaluated.

Training and Testing

The training process begins with parameter initialization, which includes the number of epochs and batch size. The dataset contains 699 instances categorized into Benign and Malignant classes. A 10-fold cross-validation scheme is employed, where the dataset is divided into ten equal parts nine assigned for training and one reserved for testing in each rotation. This process is performed across ten repetitions, and the average performance is used to measure how effectively the model performs classification. The best weights from training are saved and used in the testing phase. Model accuracy is then evaluated using a confusion matrix to assess the performance of breast cancer classification.

RESULTS AND DISCUSSION

Preprocessing

The dataset employed in this study originates from a publicly available breast cancer database, which comprises 699 data points with 9 features. Before feature selection and training using the ML architecture, data pre-processing is necessary to ensure that the model performance results are more optimal. After pre-processing, the WBC dataset consists of 458 Benign classes and 241 Malignant classes. Two features are not used in training, namely the Id feature and the Class feature. The Class feature is used as a result of breast cancer prediction, and the dataset has average values, standard deviations, and maximum values after the pre-processing stage is applied. Table 2 shows a description of the dataset after pre-processing.

Table 2. WdBC dataset description after preprocessing

Features	Features names	Feature Measurement Range		
		Mean	Standard deviation	Maximum
1	Id	-	-	-
2	Class	-	-	-
3	Clump thickness	1.760	1.492	2.484
4	Uniformity of cell size	0.449	0.506	2.484
5	Uniformity of cell shape	1.514	0.498	2.484
6	Marginal adhesion	1.438	0.477	2.484
7	Single epithelial cell size	1.582	0.349	2.484
8	Bare nuclei	1.533	0.570	2.484
9	Bland chromatin	1.605	0.407	2.484
10	Normal nucleoli	1.434	0.504	2.484
11	Mitoses	1.216	0.303	2.484

Feature Selection

Selecting informative features plays a decisive role in improving the effectiveness of classification models and learning algorithms[37]. With the help of feature selection, resulting in shorter training durations and faster prediction times. The complete results of feature selection using the filter, wrapper and embedded method are presented in Table 3.

Table 3. Number of selected features with feature selection methods

		Number of selected Features	Name of the Features
Without FS		9	All Features
Filter Methods	Chi2	5	Cell size uniformity, cell shape uniformity, adhesion at the cell margins, presence of bare nuclei, and nucleoli with normal appearance
	ANOVA	5	Cell size uniformity, cell shape uniformity, bare nuclei presence, bland-type chromatin, and nucleoli of normal morphology
	MI	5	Cell size uniformity, cell shape uniformity, size of single epithelial cells, bare nuclei presence, and chromatin with a bland pattern
Wrapper Methods	FFS	5	Cluster thickness, cell shape uniformity, cell size uniformity, presence of bare nuclei, and bland-pattern chromatin
	FBS	5	Cluster thickness, cell size uniformity, marginal-level adhesion, bare nuclei presence, and chromatin of bland appearance
Embedded Methods	LASSO	7	Cluster thickness, cell size uniformity, cell shape uniformity, marginal adhesion, presence of bare nuclei, bland chromatin characteristics, and nucleoli with normal morphology
	Ridge Embedded	9	Cluster thickness, cell size uniformity, cell shape uniformity, marginal adhesion, size of individual epithelial cells, bare nuclei presence, bland-type chromatin, nucleoli with normal appearance, and mitotic activity

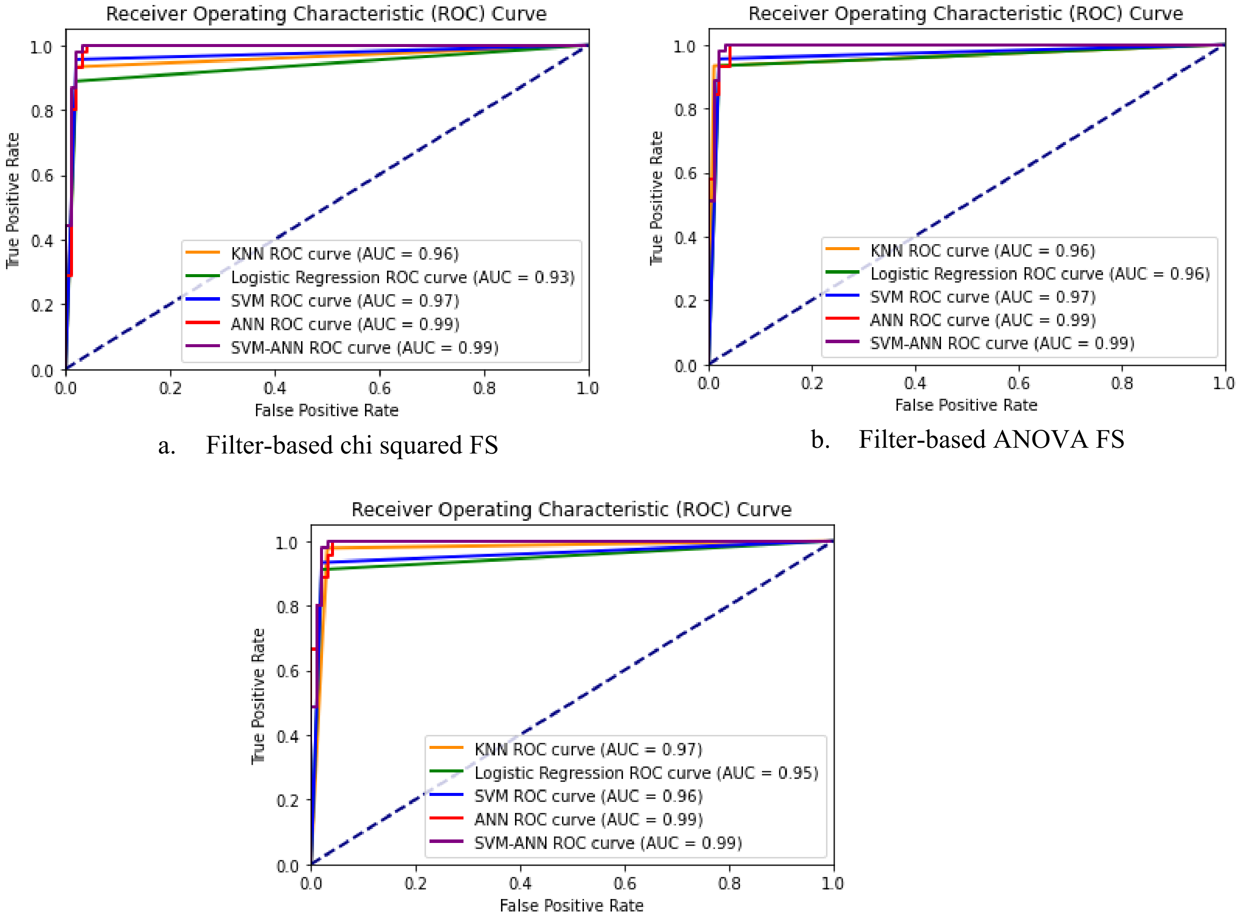
Training and Testing

This study employs several classification techniques, including KNN, Logistic Regression, SVM, ANN, and hybrid or optimized models like SVM–ANN, are employed in the classification process. Model performance is assessed using 10-fold cross-validation, in which the mean score across all folds is reported as the final evaluation metric for each algorithm. The complete results of breast cancer classification using five ML algorithms are presented in Table 4.

Table 4. Performance evaluation metrics on the WBC dataset

Feature Selection Methods	ML Algorithm	Feature Selection Methods	Acc	Pre	Re	F1
Filter Methods	KNN	Chi squared	95%	94%	94.5%	94.5%
	LR		95.88%	95.5%	95.5%	95.5%
	SVM		95.35%	94.5%	95%	95%
	ANN		95%	94.5%	95%	95%
	SVM-ANN		96%	93%	96%	95%
	KNN	ANOVA	96.24%	95.5%	96%	96%
	LR		96.06%	95.5%	95.5%	95.5%
	SVM		96.24%	95.5%	96.5%	96%
	ANN		95.89%	95.5%	96.5%	96%
	SVM-ANN		97%	92%	100%	96%
	KNN	Mutual Information	95.89%	95%	95.5%	95.5%
	LR		96.06%	95.5%	95.5%	95.5%
	SVM		96.60%	96%	96.5%	96%
	ANN		95.89%	96%	96.5%	96%
	SVM-ANN		95%	90%	96%	92%
Wrapper Methods	KNN	Forward Feature Selection	97.14%	95.5%	96%	96%
	LR		96.78%	96%	96%	96%
	SVM		96.24%	95.5%	96%	96%
	ANN		95.89%	95.5%	96%	96%
	SVM-ANN		98.57%	95.5%	96%	96%
	KNN	Backward Feature Selection	96.24%	95.5%	96%	96%
	LR		96.42%	96%	96%	96%
	SVM		96.24%	95.5%	96%	96%
	ANN		96.79%	95.5%	96%	96%
	SVM-ANN		98.57%	95.5%	96%	96%
Embedded Methods	KNN	LASSO	95.88%	95%	95.5%	95.5%
	LR		96.42%	96%	96%	96%
	SVM		96.60%	96%	96.5%	96%
	ANN		96.79%	96%	96.5%	96%
	SVM-ANN		96.43%	96%	96.5%	96%
	KNN	Ridge Embedded	96.24%	95.5%	96%	96%
	LR		96.42%	96%	96%	96%
	SVM		96.78%	96.5%	96.5%	96%
	ANN		96.96%	96.5%	96.5%	96%
	SVM-ANN		97.86%	96.5%	96.5%	96%

Testing conducted on the WDBC dataset using 7 feature selection methods showed that all classification models produced ROC curves that were well above the diagonal line with very high AUC values in the range of 0.93 to 1.00. The wrapper and filter methods consistently improve the discriminating ability of models such as KNN, SVM, ANN, and SVM-ANN with an AUC of around 0.95–0.99, while the embedded methods (Ridge and LASSO) provide the best performance, which can be seen in Figure 3, 4 and 5.

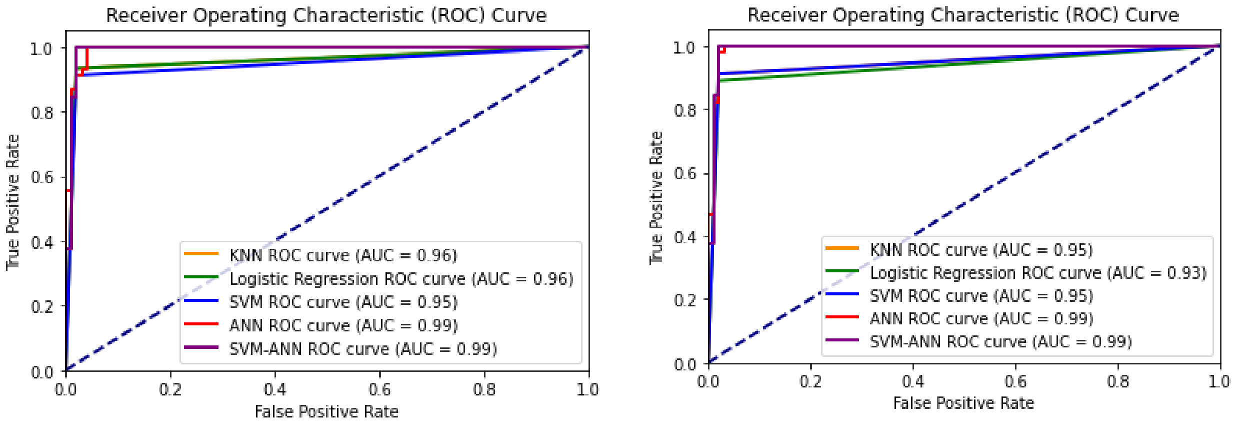


a. Filter-based chi squared FS

b. Filter-based ANOVA FS

c. Filter-based mutual information FS

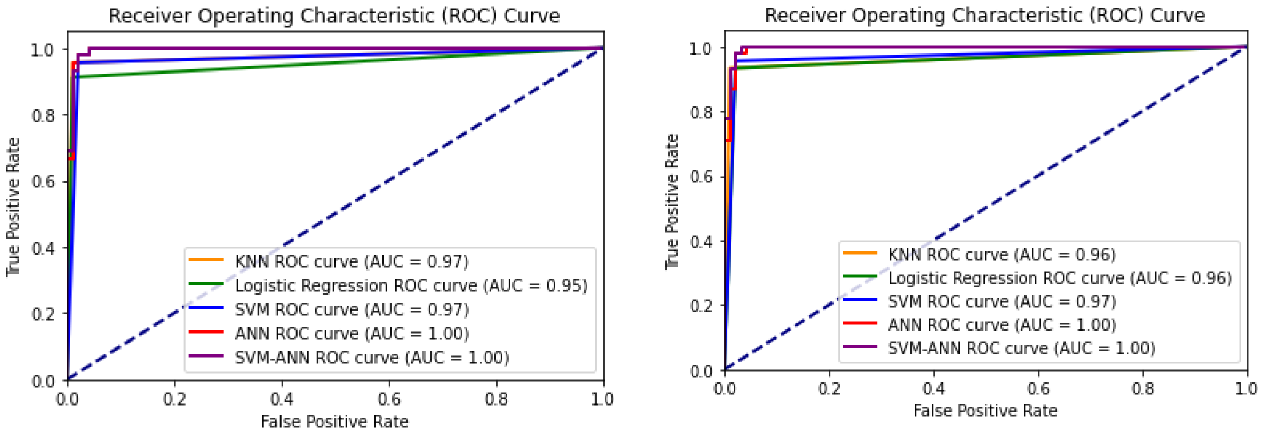
Figure 3. (a) ROC curve and AUC of filter-based chi squared FS (b) ROC curve and AUC of filter-based ANOVA FS (c) ROC curve and AUC of filter-based mutual information FS



a. Backward-based wrapper FS

b. Forward-based wrapper FS

Figure 4. (a) ROC curve and AUC of backward-based wrapper FS (b) ROC curve and AUC of forward-based wrapper FS



a. Embedded-based LASSO FS

b. Embedded-based Ridge FS

Figure 5. (a) ROC curve and AUC of embedded-based LASSO FS (b) ROC curve and AUC of embedded-based ridge FS

Discussion

Several previous studies that conducted breast cancer detection using the Wisconsin Breast Cancer dataset used several ML algorithms and applied feature selection, and combined several ML algorithms and hybrid methods, a detailed overview is presented in Table 5.

Table 5. Comparison with other studies

Methodology	Acc	Pre	Recall	F1-Score
KNN, Chi Squared Feature Selection (Mushtaq et al.) [11]	89.14%	-	89.83%	91.77%
SVM-ANN (Nathiya et al.) [15]	97%	94%	95%	94%
Logistic Regression, Feature Forward Selection (Dhanya et al.) [33]	96.49%	-	-	-
Logistic Regression (Hasan et al.) [23]	97.1%	97.1%	97.1%	-
Proposed Method (SVM-ANN, Ridge Embedded)	97.86%	96.5%	96.5%	96%

In Table 5, research conducted by Musthaq et al. [11] employed the ML KNN algorithm and feature selection using the Chi-Squared test, yielding accuracy, precision, and recall values above 91%. However, the study did not measure precision. Research conducted by Nathiya et al. [15] combined two ML algorithms, namely SVM and ANN, into a Hybrid algorithm in detecting breast cancer and obtained an average accuracy, precision, recall, and f1-score of 95% but the study did not apply feature selection. Research conducted by [33] applied the ML LR algorithm to detect breast cancer, achieving an accuracy of 96.49%. However, it did not measure other performance metrics. Research conducted by Hasan et al. [23]using the ML LR algorithm to detect breast cancer yielded results in the form of accuracy, precision, and recall of 97.1%, but did not measure the F1-score. In this study, using the SVM-ANN algorithm and the embedded feature selection method with the Ridge Embedded technique, the highest model performance results were achieved, yielding an accuracy of 97.86%, a precision of 96.5%, a recall of 96.5%, and an F1-score of 96%.

CONCLUSION

The findings indicate that the preprocessing procedures play a crucial role in resolving data inconsistencies, thereby enabling optimal model performance during training. At the feature selection stage, it is carried out to select important features and remove unimportant features to speed up the training process and obtain maximum results. An exploration of the most effective feature selection method for improving model performance was conducted, specifically using the embedded method with the Ridge Embedded technique.

At the classification stage, exploration was conducted on six ML algorithms and the hybrid SVM-ANN algorithm with Ridge Embedded feature selection, which produced the highest model performance in detecting breast cancer, with an average accuracy of 96.8%. These results indicate that the hybrid SVM-ANN algorithm is robust and capable of classifying breast cancer effectively, making it suitable for the early detection.

REFERENCES

[1] R. Widiasih, Ermiati, T. N. Jayanti, and Y. Rais, "Psychosocial Interventions for Improving the Quality of Life in Breast Cancer Survivors: A Literature Review," in *IOP Conference Series: Earth and Environmental Science*, 2019. doi: 10.1088/1755-1315/248/1/012056.

[2] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *Lancet Glob Health*, vol. 4, no. 9, 2016, doi: 10.1016/S2214-109X(16)30143-7.

[3] R. Siegel *et al.*, "Cancer treatment and survivorship statistics, 2012," *CA Cancer J Clin*, vol. 62, no. 4, 2012, doi: 10.3322/caac.21149.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, no. 6, 2018, doi: 10.3322/caac.21492.

[5] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in *2021 International Conference on Artificial Intelligence, ICAI 2021*, Institute of Electrical and Electronics Engineers Inc., Dec. 2021, pp. 97–101. doi: 10.1109/ICAI52203.2021.9445249.

[6] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," in *SIAM News*, 1990.

[7] V. Singhal, Y. Chaudhary, S. Verma, U. Agarwal, and Mr. P. Sharma, "Breast Cancer Prediction using KNN, SVM, Logistic Regression and Decision Tree," *Int J Res Appl Sci Eng Technol*, vol. 10, no. 5, 2022, doi: 10.22214/ijraset.2022.42688.

[8] M. Abdar *et al.*, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognit Lett*, vol. 132, 2020, doi: 10.1016/j.patrec.2018.11.004.

[9] L. Barracliffe, O. Arandjelović, and G. Humphris, "A pilot study of breast cancer patients: Can machine learning predict healthcare professionals' responses to patient emotions?," in *Proceedings of the 9th International Conference on Bioinformatics and Computational Biology, BICOB 2017*, 2017.

[10] C. Birkett, O. Arandjelovic, and G. Humphris, "Towards objective and reproducible study of patient-doctor interaction: Automatic text analysis based VR-CoDES annotation of consultation transcripts," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2017. doi: 10.1109/EMBC.2017.8037399.

[11] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers,Series A*, vol. 43, no. 1, 2020, doi: 10.1080/02533839.2019.1676658.

[12] E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*, 2019. doi: 10.1109/EBBT.2019.8741990.

[13] A. Purwinarko, K. Budiman, A. Widiyatmoko, F. A. Sasi, and W. Hardyanto, "Integrating C4. 5 and K-Nearest Neighbor Imputation with Relief Feature Selection for Enhancing Breast Cancer Diagnosis," *Scientific Journal of Informatics*, vol. 12, no. 1, pp. 107–118, 2025.

[14] M. S. Rahman, M. Ahmed, A. K. B. Arnob, and M. A. Niam, "KNN for Breast Cancer Prediction utilizing Wisconsin Cancer Dataset".

[15] S. Nathiya, J. Sumitha, M. Varun, S. Suganya, and G. Sathana, "SVM-ANN Optimized Algorithm for the Classification of Breast Cancer Data as Benign and Malignant," in *Proceedings - 2nd*

International Conference on Smart Technologies, Communication and Robotics 2022, STCR 2022, 2022. doi: 10.1109/STCR55312.2022.10009301.

[16] S. Tokgöz and M. Açkkar, “A Hybrid Classifier Combining SVM and CNN with Feature Selection for Predicting Breast Cancer Diagnosis,” in *Science and Engineering Congress Bildirini No: 228*, 2022, pp. 535–541.

[17] T. S. Lim, K. G. Tay, A. Huong, and X. Y. Lim, “Breast cancer diagnosis system using hybrid support vector machine-artificial neural network,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, 2021, doi: 10.11591/ijece.v11i4.pp3059-3069.

[18] B. Sahu, S. N. Mohanty, and S. K. Rout, “A Hybrid Approach for Breast Cancer Classification and Diagnosis,” *EAI Endorsed Transactions on Scalable Information Systems*, vol. 6, no. 20, 2019, doi: 10.4108/eai.19-12-2018.156086.

[19] N. Bidi and Z. Elberrichi, “Feature selection for text classification using genetic algorithms,” in *Proceedings of 2016 8th International Conference on Modelling, Identification and Control, ICMIC 2016*, 2017. doi: 10.1109/ICMIC.2016.7804223.

[20] N. C. López, M. T. García-Ordás, F. Vitelli-Storelli, P. Fernández-Navarro, C. Palazuelos, and R. Alaiz-Rodríguez, “Evaluation of feature selection techniques for breast cancer risk prediction,” *Int J Environ Res Public Health*, vol. 18, no. 20, 2021, doi: 10.3390/ijerph182010670.

[21] T. A. Assegie, R. L. Tulasi, V. Elanangai, and N. K. Kumar, “Exploring the performance of feature selection method using breast cancer dataset,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, 2022, doi: 10.11591/ijeecs.v25.i1.pp232-237.

[22] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim, “Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification,” in *Lecture Notes in Intelligent Transportation and Infrastructure*, vol. Part F1405, 2019. doi: 10.1007/978-3-030-11196-0_28.

[23] R. Hasan and A. S. M. Shafi, “Feature Selection based Breast Cancer Prediction,” *International Journal of Image, Graphics and Signal Processing*, vol. 15, no. 2, 2023, doi: 10.5815/ijigsp.2023.02.02.

[24] A. Khoirunnisa and N. G. Ramadhan, “Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy,” *JURNAL INFOTEL*, vol. 15, no. 4, 2023, doi: 10.20895/infotel.v15i4.1056.

[25] E. Akkur, F. TURK, and O. Erogul, “Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization,” *Computer Systems Science and Engineering*, vol. 45, no. 2, 2023, doi: 10.32604/csse.2023.033003.

[26] S. Farahdiba, D. Kartini, R. A. Nugroho, R. Herteno, and T. H. Saragih, “Backward Elimination for Feature Selection on Breast Cancer Classification Using Logistic Regression and Support Vector Machine Algorithms,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 4, 2023, doi: 10.22146/ijccs.88926.

[27] D. D. Pramesti, Y. Farida, D. C. R. Novitasari, and A. Teguh, “Breast Cancer Classification Based on Mammogram Images Using CNN Method with NASNet Mobile Model,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 19, no. 3.

[28] S. Karim, A. Nurhuda, and others, “Prostate Cancer Detection Using Gradient Boosting Machines Effectively,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 19, no. 3.

[29] R. Shafique *et al.*, “Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning,” *Cancers (Basel)*, vol. 15, no. 3, 2023, doi: 10.3390/cancers15030681.

[30] K. Boutahar, S. Laghmati, H. Moujahid, O. El Gannour, B. Cherradi, and A. Raihani, “Exploring Machine Learning Approaches for Breast Cancer Prediction: A Comparative Analysis with ANOVA-Based Feature Selection,” in *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IRASET60544.2024.10549284.

[31] D. Jain and V. Singh, "Diagnosis of breast cancer and diabetes using hybrid feature selection method," in *PDGC 2018 - 2018 5th International Conference on Parallel, Distributed and Grid Computing*, 2018. doi: 10.1109/PDGC.2018.8745830.

[32] N. Naveed, H. T. Madhloom, and M. S. Husain, "Breast cancer diagnosis using wrapper-based feature selection and artificial neural network," *Applied Computer Science*, vol. 17, no. 3, 2021, doi: 10.23743/acs-2021-18.

[33] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "F-test feature selection in Stacking ensemble model for breast cancer prediction," in *Procedia Computer Science*, 2020. doi: 10.1016/j.procs.2020.04.167.

[34] M. S. H. Shaon, T. Karim, M. S. Shakil, and M. Z. Hasan, "A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction," *Healthcare Analytics*, vol. 6, Dec. 2024, doi: 10.1016/j.health.2024.100353.

[35] M. Li, G. Nanda, S. S. Chhajedss, and R. Sundararajan, "Machine learning-based decision support system for early detection of breast cancer," *Indian Journal of Pharmaceutical Education and Research*, vol. 54, no. 3, 2020, doi: 10.5530/ijper.54.3s.171.

[36] S. Guo and L. Zhang, "Statistical Robustness of Empirical Risks in Machine Learning," 2023. [Online]. Available: http://jmlr.org/papers/v24/20-1039.html.

[37] *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019.