Arab Journal of Management, Banking, and Financial Studies

المجلة العربية للدراسات الإدارية والمالية والمصرفية

Using an Adaptive Linear Support Vector Machine Algorithm for Predicting the Breast Cancer

Abdulqader M. Mohsen [1,2,*], Ahmed Saleh Khaled Abdullah Alhurdi [1]

Received: 04/10/2024, Reviewed: 22/12/2024, Accepted: 12/02/2025. https://doi.org/10.59559/ajmbfs.1.1.6

Abstract: Breast cancer is the most common type of cancer and a significant contributor to the high death rates among women. The death rate increases when this condition is manually diagnosed causing delay of cancer detection since it takes several hours and requires the availability of specialists. Therefore, an automated breast cancer diagnosis has been suggested to speed up detection and stop the disease from spreading. Over the years, machine learning classification algorithms have been used to predict breast cancer. In previous studies, one of the most used algorithms is the Support Vector Machine (SVM). However, these studies have inconsistent results. This work investigates the impact of the features' selection, hyperparameters of SVM, and the mechanism of splitting data on the algorithm performance, thus, building an SVM, as a single machine learning model, that achieves a higher accuracy. The Wisconsin dataset was used to train and test this model. The experimental results showed that the performance of the model was affected by the features' selection, hyperparameters, and the mechanism of splitting data and random state values in terms of the best results and the average of the top three results. The comparison results revealed the superiority of the proposed method over the other state-of-the-art methods.

Keywords: Breast cancer prediction, Support Vector Machine, hyperparameters, machine learning.

استخدام خوارزمية خطية تكيفية لآلة متجهات الدعم في التنبؤ بسرطان الثدي

عبدالقادر محمد العبادي [1،*]، أحمد صالح خالد عبدالله الهردي [1]

الاستلام: 2024/10/04، التحكيم: 2024/12/22، القبول: 2025/02/12

الملخص: يعد سرطان الثدي أكثر أنواع السرطان شيوعًا ومساهمًا رئيسيًا في ارتفاع معدلات الوفيات بين النساء. تزداد معدلات الوفاة عندما يتم تشخيص هذه الحالة يدويًا، حيث يستغرق ذلك وقتا طويلا ويحتاج إلى متخصصين. لذلك، تم اقتراح تشخيص آلي لسرطان الثدي لتسريع عملية الكشف ومنع انتشار المرض. على مر السنين، تم استخدام خوارزميات التصنيف في تعلم الآلة للتنبؤ بسرطان الثدي. في الدراسات السابقة، كانت إحدى أكثر الخوارزميات استخدامًا هي خوارزمية آلة متجهات الدعم (SVM). ومع ذلك، كان لتلك الدراسات نتائج غير متسقة. وتبحث هذه الدراسة في تأثير اختيار الميزات، ومعلمات hyperparameters لخوارزمية SVM، وآلية تقسيم البيانات على أداء الخوارزمية. وبالتالي، تم بناء SVM كنموذج تعلم آلي فردي يحقق نتائج أعلى. وتم استخدام مجموعة بيانات Wisconsin لتدريب واختبار هذا النموذج. وقد أظهرت النتائج التجريبية أن أداء النموذج تأثر باختيار الميزات، ومعلمات hyperparameters، وآلية تقسيم البيانات وقيم الحالة العشوائية من حيث أفضل النتائج الأولى ومتوسط أفضل ثلاث نتائج. وكشفت نتائج المقارنة تفوق الطريقة المقترحة على الطرق الأخرى المستخدمة في نفس المجال.

الكلمات المفتاحية: التنبؤ بسرطان الثدي، خوارزمية آلة متجهات الدعم (SVM)، معلمات hyperparameters، تعلم الآلة.

1 Arab Academy for Management, Banking, and Financial Sciences, Aden, Yemen.
2 University of Aden, Aden, Yemen.
* Corresponding author email: a.alabadi.fcit@aden-univ.net

1 الأكاديمية العربية للعلوم الإدارية والمالية والمصرفية، عدن، اليمن.
2 جامعة عدن، عدن، اليمن.
* الباحث المراسل: a.alabadi.fcit@aden-univ.net

I. INTRODUCTION

Breast cancer, a malignant condition originating in the breast tissue, can develop unilaterally or bilaterally. It arises when cells proliferates uncontrollably, leading to tumor formation (American Cancer Society, 2025). While breast cancer is more prevalent in older women, approximately 5% of cases occur in individuals under the age of 40. These younger patients often face more aggressive forms of the disease, necessitating alternative treatment approaches. According to data from the American College of Surgeons' Cancer database (1998–2005), patients under 40 constituted a significant cohort in breast cancer studies (Sariego, 2010). Breast cancer remains a critical public health issue, with 12% of women in the United States diagnosed during their lifetime. In 2017, over 250,000 new cases were reported, underscoring the disease's widespread impact (Waks & Winer, 2019). Historically, breast cancer has been a leading cause of cancer-related mortality among women, accounting for approximately 46,300 deaths in 1993 and ranking as the second-leading cause of cancer fatalities in the U.S. (Caplan et al., 1996). The etiology of breast cancer is multifactorial, with both genetic (e.g., DNA mutations) and environmental factors contributing to its development (Hulka & Stark, 1995).

Early detection of breast cancer is crucial, as it significantly improves survival rates. In this context, data mining and machine learning techniques have emerged as powerful tools for early-stage diagnosis. Among these, classification methods such as Bayesian Classification, Decision Tree Induction, Neural Networks, and Support Vector Machines (SVM) are widely utilized (Oprea & Ti, 2014). SVM, in particular, has gained prominence due to its robust diagnostic capabilities and effectiveness in handling medical datasets. As a supervised machine learning algorithm, SVM is employed for both classification and regression tasks, making it a valuable tool for developing technologies aimed at early breast cancer detection (Rejani & Thamarai, 2009).

The motivation for applying SVM to breast cancer diagnosis stems from its numerous advantages. SVM is renowned for its robustness and efficacy in classifying medical data, making it a preferred choice for researchers (Janardhanan & Sabika, 2015). Globally, cancer remains a significant health burden, with 18.1 million new cases reported in 2020, of which 8.8 million involved women (World Cancer Research Fund, 2020). SVM operates by identifying support vector points and drawing a hyperplane between classes, maximizing the margin of separation. This approach is particularly effective in high-dimensional spaces and scenarios where the number of features exceeds the number of samples. Additionally, SVM is memory-efficient, further enhancing its practicality for large-scale medical datasets (Raj, 2022).

Despite its widespread use, studies applying SVM to distinguish between benign and malignant breast cancers have yielded inconsistent results. This inconsistency highlights the need for further investigation into the impact of hyperparameters on SVM's performance. The primary contribution of this study is to explore the influence of hyperparameters tuning on SVM's effectiveness and to develop an optimized model capable of achieving higher diagnostic accuracy. By doing so, this research aims to advance the field of machine learning in breast cancer diagnosis. Specific contributions include: (1) a comprehensive investigation to identify the optimal combination of parameters for SVM, and (2) the proposal of a novel SVM model with enhanced accuracy. Through these efforts, this study seeks to provide a reliable machine learning-based tool for the early and accurate detection of breast cancer, thereby contributing to improved patient outcomes and advancing the body of knowledge in this critical area.

The remainder of this paper is organized as follows. Section II provides a literature review of existing studies on breast cancer prediction using machine learning techniques. Section III details the methodology, including data preprocessing, feature selection, and the proposed SVM model. Section IV describes the experimental setup and evaluation metrics. Section V presents the results and discussion, comparing the proposed model with state-of-the-art methods. Finally, Section VI concludes the paper and suggests directions for future research.

II. LITERATURE REVIEW

This section reviews various studies on breast cancer diagnosis and survival prediction, focusing on the approaches employed, algorithm performance, and factors influencing predictive accuracy. While several studies achieved high classification accuracy, others reported inconsistent results, highlighting the need to explore the factors affecting algorithm performance.

A. Feature Selection in Breast Cancer Prediction

Several studies have emphasized the importance of feature selection in improving algorithm performance. Bhukya and Manchala (2022) proposed a rough set-based feature selection approach for breast cancer prediction, using algorithms such as Decision Trees (DT), k-Nearest Neighbors (KNN), Bayesian Networks (BN), Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and AdaBoost. Among these, the RF algorithm achieved the highest accuracy of 95.23%. Similarly, Saoud et al. (2019) employed feature selection techniques to enhance the accuracy of multiple algorithms, including NB, SVM, KNN, DT, LR, and ANN. Their findings showed that the NB algorithm performed best, achieving an accuracy of 97.42% when specific features were selected. These studies highlight that careful feature selection can significantly enhance the performance of machine learning models.

B. Dimensionality Reduction and Data Preprocessing Techniques

In addition to feature selection, dimensionality reduction techniques have been shown to improve predictive performance. For instance, Egwom et al. (2022) applied a Linear Discriminant Analysis–Support Vector Machine (LDA–SVM) model, achieving an accuracy of 99.20% after reducing and separating the dataset. They also compared this model with PCA-SVM and other hybrid techniques, demonstrating the effectiveness of LDA in enhancing classification accuracy. Similarly, Li and Chen (2018) evaluated machine learning methods in R, finding that Decision Trees and RF performed well with accuracies of 96.1% when preprocessing techniques were applied. These studies underline the importance of preprocessing methods such as dimensionality reduction and data normalization in improving model performance.

C. Comparative Studies of Machine Learning Algorithms

Several studies have compared different machine learning algorithms to identify the most effective models for breast cancer prediction. Aishwarja et al. (2021) explored algorithms such as RF, SVM, KNN, and BN using data from the UCI Wisconsin Breast Cancer (WBC) dataset. Dividing the data into 80% training and 20% testing, they reported that the KNN algorithm achieved the highest accuracy of 95.90%. Similarly, Showrov et al. (2019) compared SVM, BN, and ANN classifiers, with the SVM using a linear kernel achieving the highest accuracy of 96.72%. Additionally, Chaurasia et al. (2018) tested J48, Naïve Bayes (NB), and Random Projection Forest (RPF) algorithms, reporting that NB achieved an accuracy of 97.30%. These comparative studies provide valuable insights into the relative strengths and weaknesses of various algorithms. Asri et al. (2016) used a Weka tool to import a database from the site of UCI (WBC original dataset) website. Many ML methods were used, including C45, SVM, BN, and KNN. The SVM achieved a higher accuracy of 97.13%.

D. Ensemble Learning Techniques

Another avenue for improving classification accuracy is the use of ensemble techniques, which combine multiple algorithms to exploit their complementary strengths. For instance, Elnahas et al. (2019) introduced an ensemble approach that integrated SVM, RF, and ANN algorithms, achieving an accuracy of 98.50%. Similarly, Liu et al. (2019) used an intelligent classification model combining KNN, CSSVM, and BP algorithms, with the BP algorithm achieving an accuracy of 97.50%. Furthermore, Omara et al. (2017) employed an improved self-organizing map (DSOM) alongside other algorithms, obtaining an accuracy of 98.56%. These findings emphasize the potential of ensemble methods to improve prediction accuracy when properly designed and optimized.

E. The Role of Data Splitting and Hyperparameters Tuning

The way data is divided into training and testing sets also plays a critical role in algorithm performance. Priyanka et al. (2019), for example,

experimented with different data splits (e.g., 90–10% and 80–20%) and tested the KE-Sieve algorithm with varying values of k. Their results showed that the KNN algorithm achieved an accuracy of 96.35% when k=3 and a 90–10% split was used. Similarly, Hamsagayathri and Sampath (2017) analyzed breast cancer classification using Decision Tree classifiers in the Weka tool and reported that RF achieved the highest accuracy of 96.70%. These studies highlight that careful data splitting and hyperparameters tuning are essential for maximizing algorithm performance.

F. Challenges and Variability in Algorithm Performance

Despite the advancements in machine learning, significant variability exists in the reported accuracies of algorithms across studies. For instance, Shawarib et al. (2020) employed the Java Neural Network (JNN) tool with an ANN algorithm, but the model achieved a relatively low accuracy of 88.24%. This contrasts sharply with other studies where ANN-based models performed significantly better. Similarly, Bazazeh and Shubair (2016) compared RF, SVM, and NB algorithms, reporting that NB achieved an accuracy of 97.20%. Such inconsistencies suggest that algorithm performance is influenced by factors beyond the choice of the algorithm itself.

G. Factors Influencing Algorithm Performance

The variability in performance can be attributed to several factors. First, preprocessing techniques such as feature selection, dimensionality reduction, and data normalization are critical. For example, converting large numbers into a 0–1 range, handling missing values, and reducing irrelevant features can significantly impact accuracy. Second, the hyperparameters of algorithms and the tools used (e.g., Python, MATLAB, Weka) play a crucial role. Studies that optimized hyperparameters reported better performance compared to those that did not. Third, advancements in algorithms themselves, including the use of ensemble methods, has contributed to improved results. Lastly, the specific experimental setup, such as the random state used in data splitting, can also influence the results.

III. METHODOLOGY

The general architecture of the proposed model consists of four phases, which are illustrated in Figure 1. These phases are data collection, data processing, hyperparameters tuning, the proposed model (SVM), and evaluation of the model.

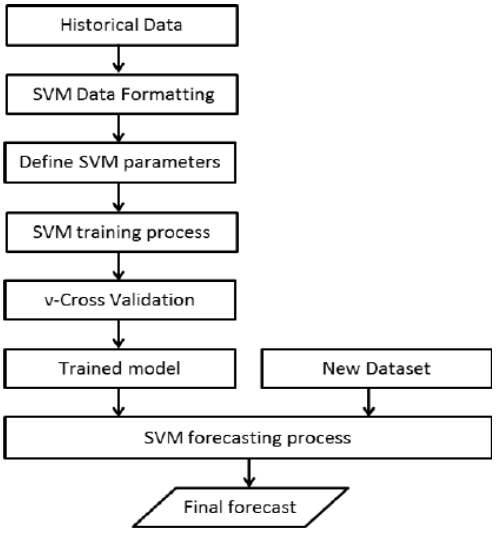

Figure 1: SVM-led Mechanism

Figure 2 illustrates these phases which are involved in building and evaluating a Support Vector Machine (SVM) model, starting with loading the dataset from the UCI repository, the breast cancer dataset, and applying data preprocessing techniques such as normalization, data conversion, and dimensionality reduction. The next stage involves choosing appropriate hyperparameters, including feature selection, train-test splitting, and setting a random state for reproducibility. The model is then built by initializing the SVM, training it on the training

data, and testing it on the testing data. Finally, the model's performance is evaluated using various metrics like accuracy, precision, recall, and F1-score.

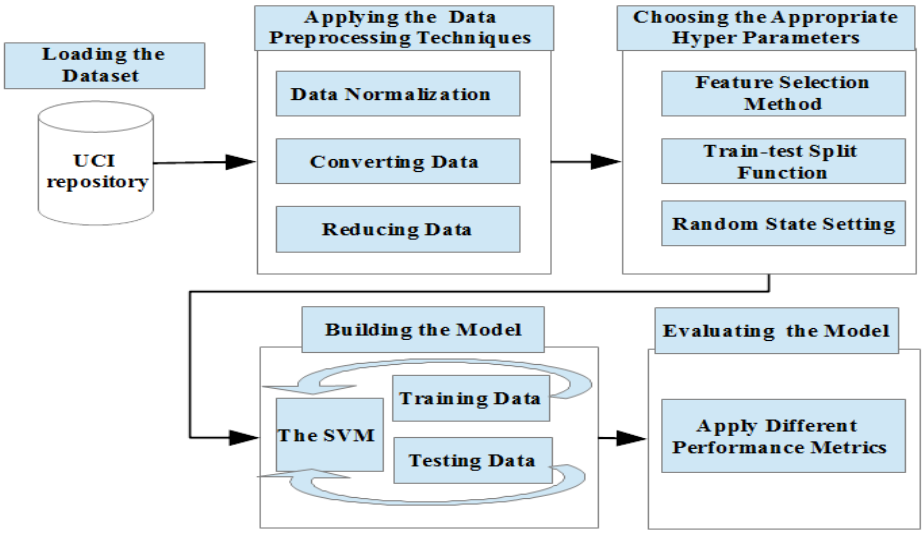

Figure 2: Proposed Model for Predicting the Breast Cancer

A. The proposed SVM

Figure 3 outlines a systematic six-step process for investigating the performance of the algorithm. In Step One, the breast cancer dataset is optimally preprocessed and represented to ensure its suitability for analysis. Step Two focuses on feature selection, where the dataset is refined by selecting from nine, eight, seven, or five features to identify the most relevant attributes for the model. Step Three involves partitioning the dataset into training and testing sets using different split ratios, specifically 90-10, 80-20, 70-30, and 60-40, to evaluate the model's performance under varying data distributions. In Step Four, random state values ranging from 1 to 10 are applied to ensure the reproducibility and robustness of the results by accounting for variability in data shuffling. Step Five documents the highest accuracy achieved across all iterations, providing a benchmark for the model's performance. Finally, Step Six calculates the average results from the top three accuracies, offering a more comprehensive and reliable measure of the algorithm's effectiveness. This structured approach ensures a thorough evaluation of the algorithm's performance while maintaining methodological rigor.

Figure 3: Algorithm's Performance Investigation Process

B. Dataset Collection

The Wisconsin breast cancer (original) dataset from the UCI Machine Learning Repository consists of 699 breast cancer cases from Wisconsin, with 458 classified as benign and 241 as malignant. This results in a distribution of 65.5% malignant and 34.5% benign cases. The dataset includes 11 integer-valued features (Dua & Graff, 2019), which are as follows: Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.

C. Data Preprocessing

Preprocessing is a crucial second phase in the proposed model, aimed at preparing the data for effective analysis and modeling. This phase consists of three key components:

1. **Reducing Data**

   This component focuses on reducing the dimensionality of the dataset while retaining essential information. Techniques such as

Principal Component Analysis (PCA) or feature selection methods can be employed to eliminate redundant or irrelevant features, thereby improving the efficiency of the model and reducing the risk of overfitting.

2. **Normalization Data**

This process involves scaling the feature values to a standard range, typically between 0 and 1. Normalization helps to eliminate any bias that may arise from varying scales of different features, ensuring that each feature contributes equally to the model's performance. This is particularly important for algorithms sensitive to the magnitude of input values, such as Support Vector Machines (SVM).

3. **Converting Data**

In this step, the data is transformed into a suitable format for analysis. This may include converting categorical variables into numerical representations, such as using one-hot encoding or label encoding. Additionally, any non-numeric data may be converted into numeric formats to facilitate processing by machine learning algorithms.

D. Appropriate Hyperparameters

The third phase is considered in the proposed model and it consists of the following elements: train test split function, feature selection, and random state.

E. The Proposed Support Vector Machine Model

This research employed both linear and kernel Support Vector Machines (SVM) to differentiate between negative and positive patient data. The SVM mechanism relies on several critical factors, which include the processed data and the configuration of hyperparameters. Key elements affecting the SVM performance are:

- **Random State:** This parameter controls the shuffling of the data, ensuring reproducibility and consistency in the results.
- **Feature Selection:** Various combinations of features were tested, specifically five, seven, eight, and nine features. This selection process is vital for optimizing the model's performance, as it influences how well the SVM can classify the data.
- **Data Splitting:** The dataset was divided into training and testing sets using different ratios, specifically 80-20, 90-10, 70-30, and 60-40. These splits are essential for evaluating the model's effectiveness and generalizability, allowing for a robust assessment of its performance on unseen data.

F. Evaluation

The evaluation of the proposed model's performance is carried out using a confusion matrix, which provides a comprehensive view of the classification results. The confusion matrix comprises four fundamental metrics: true positive, false positive, false negative, and true negative.

G. Framework for Predicting Breast Cancer

A framework has been developed using the proposed model to predict breast cancer in hospital settings, as shown in Figure 4. To facilitate accurate predictions, two key assumptions are essential for this framework. The first assumption involves preparing the data, which requires a preprocessing step to ensure that it is suitable for the proposed model. The second assumption concerns the consistency of the preprocessed data. When applying the model to different patients, the preprocessed data for each individual must be input uniformly to prevent biased outcomes. According to the proposed model, each breast cancer screening result must be classified as either positive or negative.

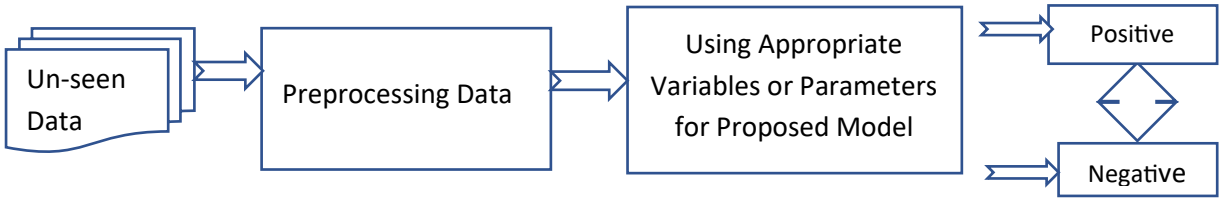

Figure 4: The Framework for Predicting Breast Cancer in Hospitals

IV. EXPERIMENTAL SETUP

This section details the overall configuration of the experiments. It describes the experimental environment, tools used, performance measures of SVM, and the parameter settings. The experiments were conducted on a machine with the following specifications: Intel (R) Core (TM) i7-4702MQ CPU @ 2.20 GHz (8 CPUs), 8.00 GB RAM, 1.0 TB hard disk drive, and Windows 10 operating system. We used Python 3.7.0 to build a single-algorithm model. Additionally, we utilized several Python libraries for predictive data analysis, including Scikit-learn, Pandas, NumPy 1.17.4, and Matplotlib. Microsoft Excel was used to organize and store datasets in tables, perform some simple preprocessing, and analyze the results. The main libraries used were NumPy, Pandas, Scikit-learn, and Matplotlib. The confusion matrix is a performance measure frequently used in classification problems with two or more class labels as output. Accuracy and F1-score are calculated using the confusion matrix. The percentage of correctly classified objects is used to calculate the classifier's accuracy. Accuracy is calculated using Equation (1) as follows:

$$\text{Accuracy} = \left(\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{True Negative}+\text{False Negative}+\text{False Positive}}\right) (1)$$

where TN and TP represent true negative and true positive, respectively. They are used to examine the correctness of the identified records as either a positive or negative class. At the same time, FN and FP denote false negatives and false positives, respectively. They are used to test the incorrectness of the identified records for the opposite class.

Precision is computed using Equation (2). Precision, referred to as confidence, is the percentage of positive and actual negative occurrences that are unmistakably positive. It demonstrates the classifier's capacity to deal with favorable findings while minimizing incorrect predictions of negative ones (Donga, 2022).

$$\text{Precision} = \left(\frac{\text{True Positive}}{\text{True Positive}+\text{FalsePositive}}\right) (2)$$

The F1-Score, which is the weighted harmonic mean of precision and recall, is calculated using Equation (3). This score accounts for both false positives and negatives.

$$\text{F1} - \text{score} = \left(2 * \frac{\text{recall} * \text{precision}}{\text{recall}+\text{precision}}\right) (3)$$

Recall, commonly referred to as sensitivity, is the frequency at which favorable predictions are correctly identified. This measure is particularly important in clinical settings, where correctly identifying a hazardous tumor is more crucial than incorrectly identifying a benign one (Donga, 2022). Recall is calculated using Equation (4):

$$\text{Recall} = \left(\frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}}\right) (4)$$

The number of features selected for various experimental investigations ranged from five to nine.

In these experiments, the data were split into training and testing sets using different configurations to assess their impact on model performance. The splitting settings included 80-20%, 90-10%, 70-30%, 60-40%, and four-fold cross-validation. The random state parameter played a crucial role in shuffling the data to optimize the algorithm's accuracy. To identify the best performance, shuffling was tested with random state values ranging from 1 to 10, although this study focused on ten specific values for random state.

In total, twenty initial experiments were conducted to evaluate and analyze the proposed combination of preprocessing steps and feature sets. Each trial involved splitting the data using the configurations of 80-20%, 90-10%, 70-30%, and 60-40%, along with random state values from 1 to 10 across the selected feature sets of five, seven, eight, and nine features. Additionally, four-fold, six-fold, and eight-fold cross-validation methods were employed to assess the performance of each configuration.

It is evident that some settings yielded high results while others produced lower outcomes. This variability can largely be attributed to the random state values used. The best results were achieved when the random state values effectively rearranged the data in a manner conducive to efficient training and testing, thereby allowing the algorithm to perform optimally. Conversely, the poorer results can be linked to less effective random state values that failed to rearrange the data appropriately, hindering the algorithm's ability to train and test efficiently. Additionally, significant disparities in the percentage of data allocated for training and testing may also have contributed to these outcomes.

V. RESULTS AND DISCUSSION

Following the investigation into a range of hyperparameters, various results were recorded, highlighting the best outcomes. The results presented in Table 1 are derived from the highest accuracy achieved. The ranking of the obtained accuracies is as follows: 100%, 99%, 98%, 97%, and 95%.

The highest accuracy of 100% was achieved with a data split of 90% for training and 10% for testing, utilizing five features and a random state of 1. This perfect accuracy was also attained with seven or eight features, while maintaining a random state of 1. The second-highest accuracy, 99%, was recorded with an 80-20% data split, using either five or nine features, both with a random state of 1. Additionally, this accuracy was achieved under a 90-10% split with nine features and a random state of 3, as well as with a 70-30% split using nine features and a random state of 9. A 99% accuracy was also noted with a 60-40% data split, employing nine features and a random state of 9. The third-highest accuracy of 98% occurred with an 80-20% split, utilizing seven features and a random state of 1. The fourth-highest accuracy of 97% was achieved across various configurations: with a 70-30% split using five or seven features and a random state of 1; an 80-20% split with eight features and a random state of 1; a 70-30% split with eight features and a random state of 1; and a 60-40% split utilizing eight features and a random state of 9. It is noteworthy that using five, seven, or eight features with a 90-10% split and a random state of 1 consistently resulted in high accuracy.

The results presented in Table 2 reflect the average of the three highest accuracies achieved. The ranking of the obtained accuracies is as follows: 99%, 98%, 97.66%, 97.33%, 97.14%, 97%, 96.66%, 96.33%, 96%, and 95%.

Table 1: Highest Accuracies Obtained Using Different Parameter Setting

Feature selection	Splits	Random state	Accuracy	F1_score
Five features	80.20	1	0.9928	0.99
	90.10	1	100	100
	70.30	1	0.9712	0.97
	60.40	1	0.9534	0.95
	KFold =4	20	0.9540	----------
Seven features	80.20	1	0.9785	0.98
	90.10	1	100	100
	70.30	1	0.9712	0.97
	60.40	9	0.9677	0.97
	KFold=4	8, 20	0.9597	---------
Eight features	80.,20	1	0.9714	0.97
	90.10	1	100	100
	70.30	9	0.9665	0.97
	60.40	9	0.9749	0.97
	KFold =4	20	0.9655	----------
Nine features	80.20	1	0.9854	0.99
	90.10	3	0.9855	0.99
	70.30	9	0.9853	0.99
	60.40	9	0.9854	0.99
	KFold =8	10	0.9763	----------

The highest accuracy of 99% was achieved with an 80-20% data split for training and testing, utilizing nine features and random state values of 1, 8, and 9. The second-highest accuracy of 98% was recorded with a 90-10% split, again using nine features, with random states of 1, 3, and 5. This accuracy was also attained with a 70-30% split using nine features and random states of 6, 9, and 10. Additionally, a 98% accuracy was noted with a 90-10% split using seven features and

random states of 1, 4, and 10, as well as with eight features under the same split and random states of 1 and 4.

The third-highest accuracy of 97.66% was achieved with a 60-40% split using nine features and random states of 6, 7, and 9, and also with a 90-10% split using five features and random states of 1, 6, and 10. The fourth-highest accuracy of 97.33% was recorded with an 80-20% split using seven features and random states of 1, 9, and 10. The fifth-highest accuracy of 97.14% was achieved with an 80-20% split, utilizing eight features and random states of 1, 9, and 10.

A sixth-highest accuracy of 97% was reached with a 70-30% split using eight features and random states of 4, 9, and 10, as well as with an 80-20% split using five features and random states of 1, 9, and 10. The seventh-highest accuracy of 96.66% was obtained with a 60-40% split using eight features and random states of 3, 9, and 10. The eighth-highest accuracy of 96.33% was achieved with both a 70-30% split using seven features and random states of 1, 5, and 9, and a 60-40% split with seven features and random states of 1, 6, and 9. The ninth-highest accuracy of 96% was achieved with a 70-30% split using five features and random states of 1, 5, and 10. Finally, the tenth-highest accuracy of 95% was recorded with a 60-40% split, utilizing five features and random states of 1, 3, and 9.

These results indicate that a significant separation between training and testing data, coupled with random state values of 1 and 10, serves as optimal hyperparameters for maximizing accuracy. Moreover, the selection of features ranging from five to nine played a crucial role in influencing the model's performance.

Table 2: Higher Results Obtained Based on the Average for the Top Three Results

Feature selection	Split	Random state	Accuracy
Five features	80.20	1,9,10	0.97
	90.10	1,6,10	0.9766
	70.30	1,5,10	0.96
	60.40	1,3,9	0.95
	KFold =4	20	0.9540
Seven features	80.20	1,9,10	0.9733
	90.10	1,4,10	0.98
	70.30	1,5,9	0.9633
	60.40	1,6,9	0.9633
	KFold=4	8, 20	0.9597
Eight features	80.20	1,9,10	0.9714
	90.10	1,4,10	0.98
	70.30	4,9,10	0.97
	60.40	3,9,10	0.9666
	KFold =4	20	0.9655
Nine features	80.20	1,8,9	0.99
	90.10	1,3,5	0.98
	70.30	6, 9,10	0.98
	60.40	6,7,9	0.9766
	KFold =8	10	0.9763

The results presented in Figure 5 and Table 3 demonstrate the high accuracy achieved based on the top results. Notably, the algorithm attained its best accuracy of 100% with a data split of 90% for training and 10% for testing, utilizing a random state value of 1 and selecting five, seven, or eight features. The second-best accuracy of 99.28% was achieved with an 80-20% split, a random state of 1, and five features. Additionally, a 90-10% split with nine features and a random state of 3 yielded an accuracy of 99%.

Table 3: Summary of the Highest Obtained Results in Table 1

Features selection	Splitting data	Random state	Accuracy
Five features	80-20%	1	99.28 %
Five features	90-10%	1	100 %
Seven features	90-10%	1	100 %
Eight features	90-10%	1	100 %
Nine features	90-10	3	99

It is important to highlight that using five, seven, and eight features with a 90-10% split and a random state of 1 consistently resulted in higher accuracy. The random state value of 1, combined with a significant separation between training and testing data, represents the optimal hyperparameters configuration for achieving superior accuracy. Moreover, the selection of features ranging from five to nine significantly influenced the model's performance.

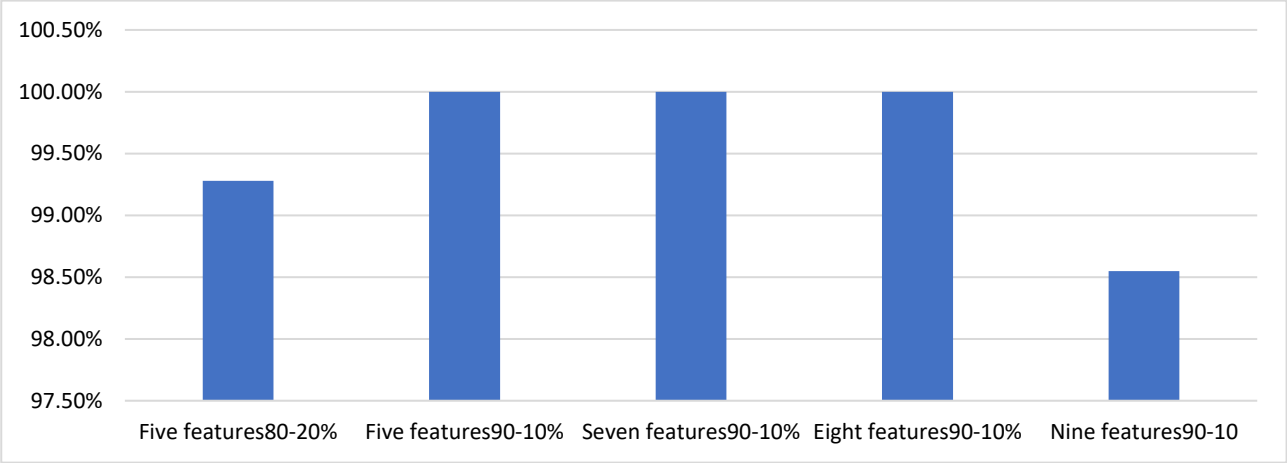

Figure 5: Best Hyperparameters Obtained the Higher Results Based on the Top One Results

Finally, Table 4 provides a summary of the highest results obtained from Table 2. This table consolidates key findings, highlighting the top performance metrics achieved in the study.

Table 4: Summary of the Highest Results Obtained in Table 2

Features selection	Splitting data	Random state	Accuracy
Seven features	90-10%	1,4,10	0.98
Eight features	90-10%	1,4,10	0.98
Nine features	80-20	1,8,9	0.99
Nine features	70-30	6,9,10	0.98
Nine features	90-10	1,3,5	0.98

The results presented in Figure 6 and Table 4 demonstrate the high accuracy achieved based on the average of the top three results. The algorithm attained its highest accuracy ranking of 99%, followed by a ranking of 98%. The top rank of 99% was achieved with an 80-20% data split, utilizing nine features and random state values of 1, 8, and 9. The second rank of 98% was recorded with a 90-10% split, using seven features and random state values of 1, 4, and 10. This same accuracy was also reached with an 80-20% split featuring eight features and random states of 1, 4, and 10, as well as with a 70-30% split using nine features and random states of 6, 9, and 10. Additionally, a 90-10% split with nine features and random states of 1, 3, and 5 also yielded an accuracy of 98%.

These results indicate that a significant separation between training and testing data, coupled with random state values of 1 and 10, serve as optimal hyperparameters for maximizing accuracy. Moreover, the selection of features ranging from five to nine played a crucial role in influencing the model's performance.

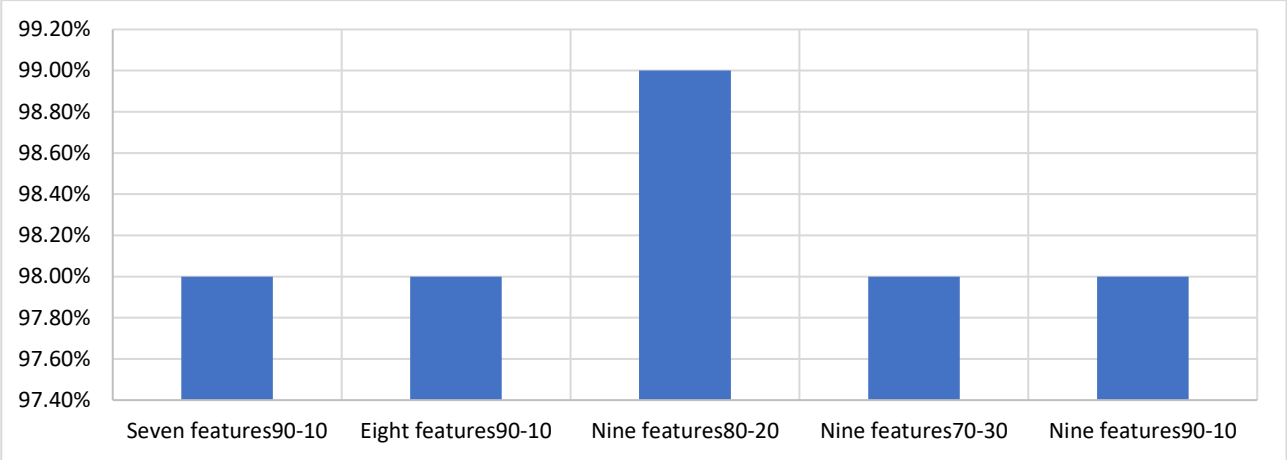

Figure 6: Best Hyperparameters to Achieve Higher Results Based on the Average for the Top Three Results

To further evaluate the classifiers' performance, the ROC curve was employed to assess their effectiveness. Figure 7 illustrates the classifier's performance across various settings. The ROC values indicate that the classifier performs exceptionally well, with values exceeding 90%. Notably, the SVM classifier achieved an ROC of 100% in several configurations: using five features with an 80-20% split, five features with a 90-10% split, seven features with a 90-10% split, eight features with a 90-10% split, and nine features with a 90-10% split.

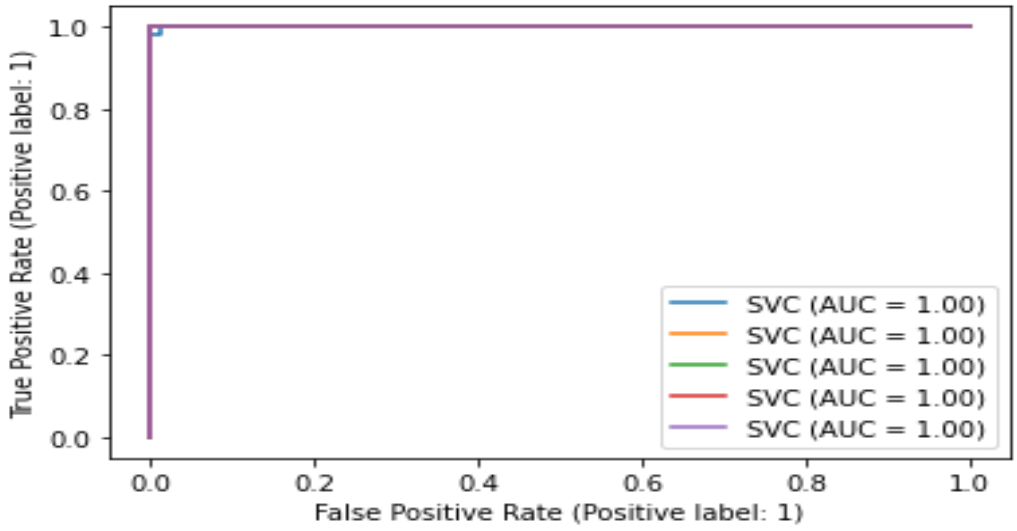

Figure 7: ROC Score of the Classifier on the Dataset

Table 5 presents a comparison of accuracy with existing methods on the breast cancer dataset. Asri et al. (2016) reported an accuracy of 97.13% using an SVM, while Kumari and Singh (2018) achieved 97.38% with the same algorithm. Liu et al. (2019) recorded an accuracy of 95.7%, and Aishwarja et al. (2021) reported 94.50%. Additionally, Bazazeh and Shubair (2016) obtained an accuracy of 97%, and Showrov et al. (2019) achieved 96.72%. Saoud et al. (2019) reported an accuracy of 97.28%, and Gbenga et al. (2017) achieved 97.7%. Lastly, Egwom et al. (2022) recorded an accuracy of 97.8%.

Table 5: Comparison to the results of previous studies with those of the proposed model

Method	Acc.	Model	Features	Dataset
proposed work with five features splitting data (90-10)	100%	SVM	5-Features	WBCO
proposed work with seven features splitting data (90-10)	100%	SVM	7-Features	WBCO
proposed work with eight features splitting data (90-10)	100%	SVM	8-Features	WBCO
proposed work with five features % splitting data (80-20)	99.28%	SVM	5-Features	WBCO
proposed work with 7 features, splitting data 90-10	98%	SVM	7-Features	WBCO
proposed work with 8 features, splitting data 90-10	98%	SVM	8-Features	WBCO
proposed work with 9 features, splitting data 80-20	99%	SVM	9-Features	WBCO
proposed work with 9 features, splitting data 90-10	98%	SVM	9-Features	WBCO
proposed work with 9 features, splitting data 70-30	98%	SVM	9-Features	WBCO
Asri et al. (2016)	97.13	SVM	9-Features	WBCO
Kumari and Singh (2018)	97.38%	SVM	9-Features	WBCO
Liu et al. (2019)	95.7%	SVM	9-Features	WBCO
Aishwarja et al. (2021)	94.50%	SVM	9-Features	WBCO
Bazazeh and Shubair (2016)	97%	SVM	9-Features	WBCO
Showrov et al. (2019)	96.72%	SVM	9-Features	WBCO
Saoud et al. (2019)	97.2818	SVM	7-Features	WBCO
Gbenga et al. (2017)	97.7%	SVM	9-Features	WBCO
Bhukya and Manchala (2022)	94.23%	SVM	9-Features	WBCO
Omara et al. (2017)	95.70%	SVM	9-Features	WBCO
Elnahas et al. (2019)	97.2%	SVM	9-Features	WBCO
Egwom et al. (2022)	97.8%	SVM	9-Features	WBCO

In contrast, the proposed model outperformed these results while utilizing the same SVM algorithm on the breast cancer dataset. This improvement can be ascribed to effective data partitioning, optimal random state values, and appropriate feature selection, all of which significantly enhanced the model's performance.

Overall, the results indicate that employing suitable hyperparameters with the classifier leads to improved classification accuracy. The proposed model demonstrates superiority in classifier accuracy, attributed to effective data partitioning, random state values, and feature selection, which collectively enhanced its performance.

VI. CONCLUSION

This study assessed the algorithm's performance with various hyperparameters configurations, finding that both the top result and the average of the top 3 results were influenced by these settings. The model achieved a maximum accuracy of 100% with a 90-10% data split and a combination of five, seven, and eight features. A second-highest accuracy of 99.28% was recorded with an 80-20% split and five features. Further configurations showed that different combinations of features and random states values, particularly with a 90-10% split, consistently led to high performance. Overall, the choice of hyperparameters significantly affects the algorithm's effectiveness.

Future research could explore additional hyperparameters, alternative algorithms, and data augmentation techniques. Investigating feature selection methods and applying the model to real-world datasets may enhance its robustness and practical applicability. In addition, future research holds significant potential to expand on this work by exploring comparisons with other machine learning algorithms, thereby placing our findings within the broader landscape of classification techniques.

REFERENCES

Aishwarja, A. I., Eva, N. J., Mushtary, S., Tasnim, Z., Khan, N. I., & Islam, M. N. (2021). Exploring the machine learning algorithms to find the best features for predicting the breast cancer and its recurrence. In P. Vasant, I. Zelinka, & G.W. Weber (Eds.), *Intelligent computing and optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)* (pp. 546-558). Springer International Publishing. https://doi.org/10.1007/978-3-030-68154-8_48

American Cancer Society. (2025, January 22). *Key statistics for breast cancer*. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICEDSA.2016.7818560

Bhukya, H., & Manchala, S. (2022). RoughSet based Feature Selection for Prediction of Breast Cancer, 29 June 2022, *PREPRINT (Version 1) available at Research Square* https://doi.org/10.21203/rs.3.rs-1542645/v1

Caplan, L. S., Helzlsouer, K. J., Shapiro, S., Wesley, M. N., & Edwards, B. K. (1996). Reasons for delay in breast cancer diagnosis. *Preventive Medicine*, *25*(2), 218-224. https://doi.org/10.1006/pmed.1996.0049

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119-126. https://doi.org/10.1177/1748301818756225

Donga, H. G. V. K. (2022). *Comparing machine learning models for diagnosis of breast cancer* [Published bachelor's thesis, Blekinge Institute of Technology]. Diva Portal. https://www.diva-portal.org/smash/get/diva2:1679145/FULLTEXT02

Dua, D., & Graff, C. (2019). *UC Irvine Machine Learning Repository*. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu

Egwom, O. J., Hassan, M., Tanimu, J. J., Hamada, M., & Ogar, O. M. (2022). An LDA–SVM machine learning model for breast cancer classification. *BioMedInformatics*, *2*(3), 345-358. https://doi.org/10.3390/biomedinformatics2030022

Elnahas, M., Hussein, M., & Keshk, A. (2019). Artificial neural network as ensemble technique fuser for improving classification accuracy. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)* (pp. 174-179). IEEE. https://doi.org/10.1109/ICICIS46948.2019.9014791

Gbenga, D. E., Christopher, N., Yetunde, D. C., & Maiduguri, N. (2017). Performance comparison of machine learning techniques for breast cancer detection. *Nova Journal of Engineering and Applied Sciences*, *6*(1), 1-8.

Hamsagayathri, P., & Sampath, P. (2017). Performance analysis of breast cancer classification using decision tree classifiers. *International Journal of Current Pharm Research*, *9*(2), 19-25. https://doi.org/10.22159/ijcpr.2017v9i2.17383

Hulka, B. S., & Stark, A. T. (1995). Breast cancer: Cause and prevention. *The Lancet*, *346*(8979), 883-887. https://doi.org/10.1016/S0140-6736(95)92713-1

Janardhanan, P., & Sabika, F. (2015). Effectiveness of support vector machines in medical data mining. *Journal of communications software and systems*, *11*(1), 25-30. https://doi.org/10.24138/jcomss.v11i1.114

Kumari, M., & Singh, V. (2018). Breast cancer prediction system. *Procedia Computer Science*, *132*, 371-376. https://doi.org/10.1016/j.procs.2018.05.197

Li, Y., & Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. *Applied and Computational Mathematics*, *7*(4), 212-216. https://doi.org/10.11648/j.acm.20180704.15

Liu, N., Qi, E. S., Xu, M., Gao, B., & Liu, G. Q. (2019). A novel intelligent classification model for breast cancer diagnosis. *Information Processing & Management*, *56*(3), 609-623. https://doi.org/10.1016/j.ipm.2018.10.014

Omara, H., Lazaar, M., & Tabii, Y. (2017). Classification of breast cancer with improved self-organizing maps. In *BDCA'17: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications* (Article No.: 73). Association for Computing Machinery. https://doi.org/10.1145/3090354.3090429

Oprea, C., & Ti, Ş. (2014). Performance evaluation of the data mining classification methods. *Information Society and Sustainable Development, 1*, 249-253.

Priyanka, G., Sahoo, P. K., Rohith, V., & Eswaran, K. (2019). Breast cancer prediction system using KE Sieve algorithm. *International Journal of Scientific & Engineering Research*, *10*(1), 19-21.

Raj, A. (2022, Mar 30). Everything about support vector classification — Above and beyond: A comprehensive read on support vector classification. https://medium.com/towards-data-science/everything-about-svm-classification-above-and-beyond-cc665bfd993e#:~:text=Support%20Vector%20Machines%20or%20SVMs,in%20an%20n%2Ddimensional%20space

Rejani, Y. I., & Thamarai, S. (2009). Early detection of breast cancer using SVM classifier technique. *International Journal on Computer Science and Engineering, 1*(3), 127-130. https://doi.org/10.48550/arXiv.0912.2314

Saoud, H., Ghadi, A., Ghailani, M., & Abdelhakim, B. A. (2019). Using feature selection techniques to improve the accuracy of breast cancer classification. In M. Ben Ahmed, A. A. Boudhir, & A. Younes (Eds.), *Innovations in smart cities applications edition 2: The proceedings of the third International Conference on Smart City Applications* (pp. 307-315). Springer International Publishing. https://doi.org/10.1007/978-3-030-11196-0_28

Sariego, J. (2010). Breast cancer in the young patient. *The American Surgeon*, *76*(12), 1397-1400. https://doi.org/10.1177/000313481007601226

Shawarib, M. Z. A., Latif, A. E. A., Al-Zatmah, B. E. E., & Abu-Naser, S. S. (2020). Breast cancer diagnosis and survival prediction using JNN. *International Journal of Engineering and Information Systems, 4*(10), 23-30.

Showrov, M. I. H., Islam, M. T., Hossain, M. D., & Ahmed, M. S. (2019). Performance comparison of three classifiers for the classification of breast cancer dataset. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-5). IEEE. https://doi.org/10.1109/EICT48899.2019.9068816

Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: A review. *Jama, 321*(3), 288-300. https://doi.org/10.1001/jama.2018.19323

World Cancer Research Fund. (2020). *Worldwide cancer data: Global cancer statistics for the most common cancers in the world.* https://www.wcrf.org/preventing-cancer/cancer-statistics/worldwide-cancer-data

To cite this article...

Mohsen, A. M., & Alhurdi, A. S. K. A. (2025). Using an adaptive linear support vector machine algorithm for predicting the breast cancer.. *Arab Journal of Management, Banking, and Financial Studies, 1*(1), 90-103. https://doi.org/10.59559/ajmbfs.1.1.6