Evaluative Comparison of Machine Learning Algorithms for Precision Diagnosis in Breast Cancer

Jiarui Li *

School of Informatics, College of Science and Engineering, University of Edinburgh, Edinburgh, EH8 9YL, UK

* Corresponding Author Email: J.Li-387@sms.ed.ac.uk

Abstract. Breast cancer remains a prominent issue in worldwide public health, exhibiting a gender disparity that primarily impacts women. This study systematically evaluates the diagnostic capabilities of various machine learning algorithms in predicting breast cancer recurrences. Utilising a dataset of 569 data points, the algorithms scrutinised include Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), XGBoost (XGB), Logistic Regression (LR), and K-Nearest Neighbours (KNN). Principal Component Analysis (PCA) was applied and employed with the algorithmic evaluation for selecting features and reducing dimensionality. The study utilised multiple evaluative metrics, focusing on Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. The findings suggest that Logistic Regression and Support Vector Machines performed better than the other algorithms. Specifically, Logistic Regression achieved an AUC value of 99.77%, and Support Vector Machines achieved an AUC value of 99.74%. Additionally, these algorithms demonstrated an accuracy rate of 97.37%, precision of 97.62%, recall of 95.35%, F1 score of 96.47%, and Cohen's Kappa coefficient of 94.37%, consistent. The study suggests potential avenues for further investigation into the utility of machine learning algorithms and dimensionality reduction techniques in diagnosing breast cancer recurrence. These preliminary findings have the potential to make a valuable contribution to the current discourse around the use of machine learning technologies within healthcare environments.

Keywords: Breast cancer; machine learning; support vector machine; logistic regression; area under the curve values.

1. Introduction

Breast cancer, a prominent contributor to cancer-related mortality in women, is a significant worldwide health issue requiring timely and precise identification to enhance treatment efficacy [1]. In 2020, the World Health Organisation reported 2.3 million breast cancer cases worldwide among women, resulting in 685,000 fatalities [2]. Traditionally, the diagnosis of breast tumours, including benign and malignant cases, has depended on manual examination, mammography, ultrasound imaging, and subsequent histopathological analyses [3]. While typical traditional procedures have shown a certain degree of efficacy, they often require significant human effort, may include subjectivity, and might provide erroneous negative or positive results. These factors introduce complexities into the diagnostic procedure and affect patient outcomes.

The emergence of computational techniques, especially machine learning (ML) and artificial intelligence (AI), has introduced a transformative approach to disease diagnosis. Machine learning, a subset of AI, specialises in developing algorithms to learn from and make data-based decisions. In recent decades, there has been a notable increase in the use of ML algorithms within the medical domain [4]. One potential advantage linked to machine learning methodologies is the ability to provide precise and prompt diagnostics for various malignancies, such as breast cancer.

Artificial Neural Networks (ANN) is a prevalent machine learning algorithm used for diagnostic purposes. These networks emulate the linked structure of neurons in the human brain and have enhanced capabilities to handle complex datasets [5]. Support Vector Machines (SVM) are recognised for their effectiveness in high-dimensional spaces, and Decision Trees (DT) offer simple, interpretable classification rules. Moreover, ensemble methods like Random Forests (RF) and gradient boosting algorithms such as XGBoost (XGB) have emerged, leveraging the power of

multiple learners for enhanced performance. Traditional algorithms such as Logistic Regression (LR) offer probabilistic insights, while K-Nearest Neighbours (KNN) gives intuitive, instance-based reasoning.

This study employs several machines learning approaches, including ANN, SVM, LR and others, to further investigate the benefits of machine learning in predictive diagnoses for breast cancer. This research aims to assess and investigate the precision of several machine learning algorithms in discriminating between benign and malignant breast cancers. This study aims to determine the model that demonstrates the highest diagnostic precision. The primary aim is to equip doctors and clinical teams with enhanced resources that might potentially facilitate the early diagnosis of breast cancer in patients and enable the timely administration of suitable treatment treatments.

2. Methodology

2.1. Data Description

This paper uses the Breast Cancer Wisconsin (Diagnostic) Database. The public may access the dataset via the UCI Irvine Machine Learning Repository [6]. The dataset comprises diagnostic data related to breast cancer, featuring 569 observations with 33 attributes. Each case encapsulates a unique diagnostic profile, highlighting the data's wide range and comprehensive nature. The main characteristic, diagnosis, is used to determine the malignancy status, categorising it as either 'Benign' (B) or 'Malignant' (M). At the same time, the outcome variable comprises 357 instances of 'B' and 212 instances of 'M'. The following characteristics explore the complex intricacies of cell nuclei features derived from the computational analysis of a digitised picture obtained from a breast mass's fine needle aspirate (FNA). These metrics include radius_mean, representing distances from the centre to various places on the perimeter, and texture_mean, denoting the values of grey-scale standard deviation. Additional notable characteristics include perimeter_mean, area_mean, and more intricate metrics such as concavity_mean and concave points_mean. When used together, these measures provide a comprehensive understanding of tumour features.

2.2. Flow Chart of Data Processing

This study compares multiple breast cancer screening algorithms to identify the most accurate. Starting with data description, followed by algorithm selection, pre-processing includes data cleansing, identify correlation, feature scaling, and data splitting. The final steps are algorithm evaluation and result, as shown in Figure 1. Detailed steps follow.

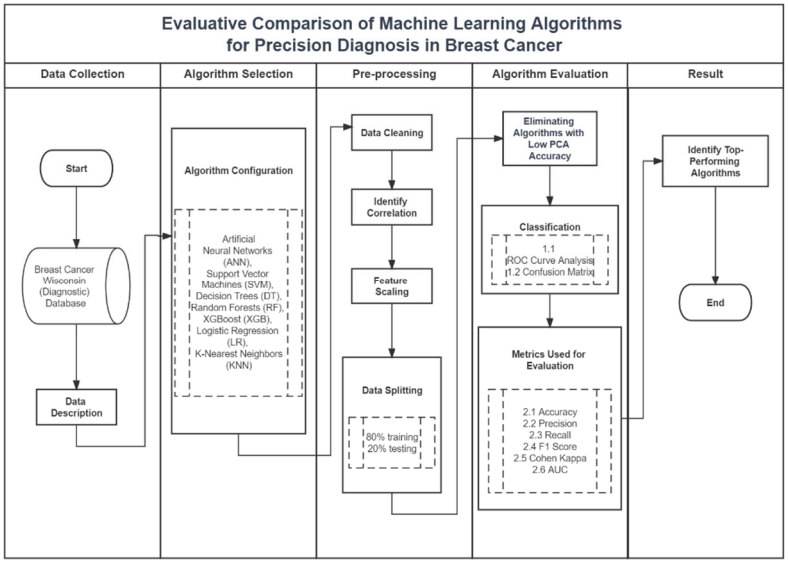

Fig. 1 Flow chart of the whole study

2.2.1 Algorithm configuration

This study explores the use of several machine learning algorithms, including ANN, SVM, DT, RF, XGB, LR, and KNN.

ANN is a computational algorithm that draws inspiration from the intricate structure and functional mechanisms seen in the neural networks of the human brain. ANN, composed of interconnected "neurons," can analyse data and identify patterns, making them indispensable in applications such as image and speech recognition, medical diagnosis, and financial forecasting. By altering connections (like brain synapses) depending on input data, ANN can learn and make autonomous judgements. Various architectures, including feedforward, recurrent, and convolutional, may be used for various applications [7].

SVM is a computational algorithm that classifies objects based on its ability to learn from sample datasets [8]. Gene expression profiles from various samples, such as tumours, are often used in biomedicine to facilitate the classification process, hence assisting in the diagnosis and prognosis of diseases. SVMs can also identify patterns across diverse data, including many domains such as credit card transactions and handwritten numerical characters [9].

DT are versatile algorithms that may be effectively used for classification and regression applications. The operational procedure involves splitting data based on feature values until homogeneous subsets are formed [10].

RF improves the performance of DT by constructing multiple trees and aggregating their outputs, resulting in enhanced accuracy and better control of overfitting [11].

XGB is a highly optimised distributed gradient boosting library developed to prioritise efficiency, flexibility, and portability. The algorithm is popular in machine learning due to its exceptional speed and performance across competitions and practical implementations. XGB employs a gradient boosting algorithm to construct algorithms iteratively, optimising for accuracy and handling various data types [12].

LR is a statistical algorithm used for predicting binary outcomes. Simulating the dependent variable's log odds calculates the chance of an event happening depending on one or more independent variables [13].

KNN is a non-parametric algorithm for learning that is often used for classification and regression problems. The classification of an unknown instance is determined by assigning it the label of the majority class among its 'k' closest instances in the training dataset [14].

2.2.2 Data cleaning

The dataset underwent pre-processing techniques to ensure its suitability for machine learning applications. The 'id' column was omitted from the dataset since it is essential to exclude all identifiers before doing the analysis. The last characteristic, "Unnamed:32," seems extraneous and without a clear purpose inside the dataset. The dataset contains missing values (NaN) that must be removed before data analysis. All columns were converted into a numeric format to satisfy the requirements of the algorithms used in this study.

2.2.3 Identify correlation.

The diagnosis outcomes are initially transformed from characters to numerical values to identify correlation. Specifically, 'M' is assigned a value of 1, while 'B' is 0. This transformation arises from acknowledging that malignant cancers often need more extensive medical intervention than benign tumours. The main objective of this procedure is to pinpoint the best algorithm for predicting benign and malignant breast cancers, emphasising the connection between the diagnosis and the 30 features. As illustrated in Figure 2, 15 features show a correlation exceeding 0.5, while only five correlate 0.1. This observation implies that a significant portion of the features should be considered throughout the study.

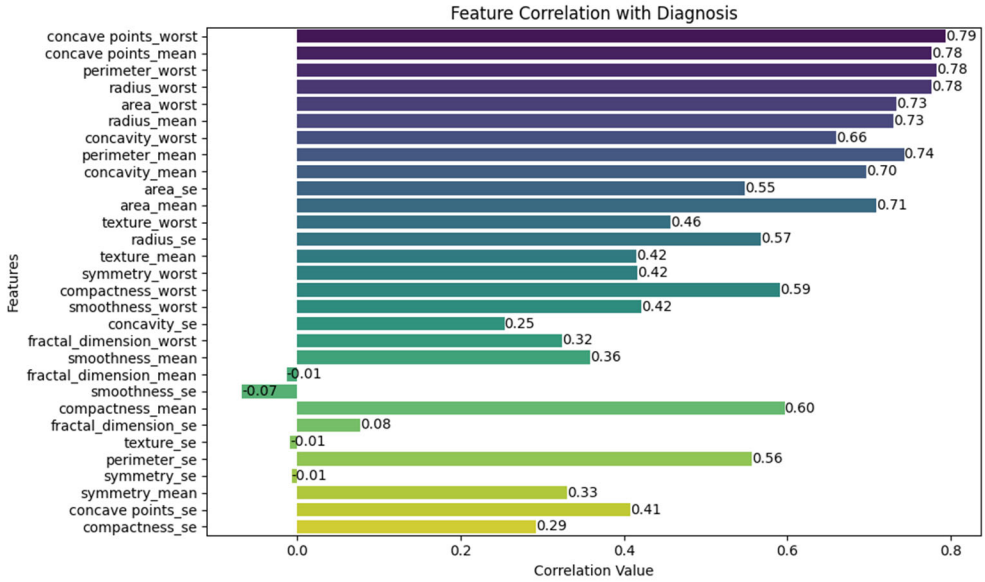

Fig. 2 30 features' correlation with diagnosis

2.2.4 Feature scaling

Given the sensitivity of algorithms such as SVM, KNN, and ANN to feature sizes, utilising Principal Component Analysis (PCA) is advantageous. PCA is a technique that converts a potentially correlated feature set into linearly uncorrelated variables, known as principal components [15]. Utilising the approach mentioned above, our research used PCA with a variance threshold of 95%, resulting in the extraction of 10 principal components. Figure 3(a) shows the variation explained by each primary component. It is evident that the early few main components, particularly the first 10, provide the most contribution to the variance of the data. This suggests that these components effectively capture the essential information contained within the data. From the cumulative explained variance curve in Fig.3(b), it can be observed that ten principal components already account for 95% of the data's variance, suggesting that the chosen dimensionality reduction has successfully preserved a significant portion of the information inherent in the original dataset.

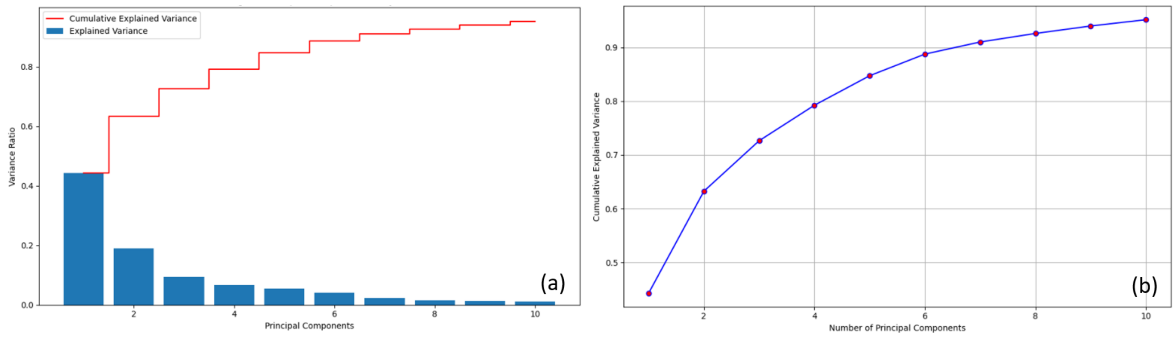

Fig. 3 PCA results of (a) line chart and (b) scree plot.

2.2.5 Data splitting

Once the features are scaled and the dimensionality is reduced, the dataset is split into several subsets for training and testing purposes. This approach objectively evaluates the algorithm's efficacy by guaranteeing that it has not just relied on memorising the training data. As mentioned, this study mainly employed the Train-test-split method, allocating 80% of the dataset for training and the remaining 20% for testing.

3. Results

3.1. Result of PCA

Table 1. Comparison of Algorithm Accuracy Using Original and PCA-Reduced Data

Algorithm	Original Data Accuracy	PCA Data Accuracy
SVM	0.935673	0.935673
LR	0.988304*	0.953216
ANN	0.953216	0.959064*
RF	0.959064	0.953216
DT	0.918129	0.923977
KNN	0.953216	0.941520
XGB	0.947368	0.953216

* Means the optimum value of Original Data Accuracy or PCA Data Accuracy among 7 algorithms

The observed consistency implies that the SVM's efficacy was preserved even when a reduced number of primary components were used for dimensionality reduction by PCA. In contrast, as shown in Table1, LR had a notable accuracy rate of 98.83% when applied to the original dataset. However, its performance exhibited a fall when applied to the dataset that had undergone PCA transformation. This observation suggests that PCA excluded key essential elements from the initial dataset. It is worth noticing that, as visualised in Figure 4, ANN exhibited improved performance after using PCA on the data. This upward trend indicates that simplifying the PCA data may have facilitated the training process for ANN. Both the RF and KNN algorithms experienced a slight decrease in performance when applied to the PCA-transformed data. This might indicate alterations in the data spaces or the possible loss of complex feature relationships. Notably, DT and XGB witnessed a boost in their performance metrics using PCA, highlighting their capacity to handle data with reduced dimensions effectively.

In general, the performance of the algorithms after applying PCA demonstrates the successful preservation of important information from the initial dataset. This highlights the effectiveness of PCA in improving computing efficiency while maintaining the performance of the algorithms.

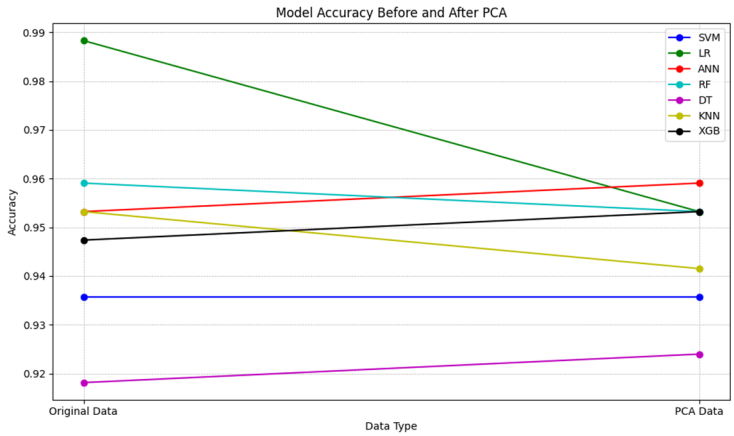

Fig. 4 Trends of each algorithm before and after PCA

3.2. Algorithm Evaluation

3.2.1 Confusion matrices

Consistent with prior research supporting SVM as the preferred algorithm for breast cancer detection, SVM was retained for further analysis despite other algorithms like DT and KNN falling short of the 95% PCA data accuracy threshold. Upon analysing Figure 5, it can be seen that both SVM and LR demonstrated zero False Positives (FP) and 72 True Negatives (TN). This indicates their proficiency in accurately identifying benign cases. Nonetheless, SVM's True Positives (TP) were

lower than LR. Specifically, LR attained an Actual Positive (AP) rate of 40, suggesting it may be slightly advantageous in identifying malignant cases.

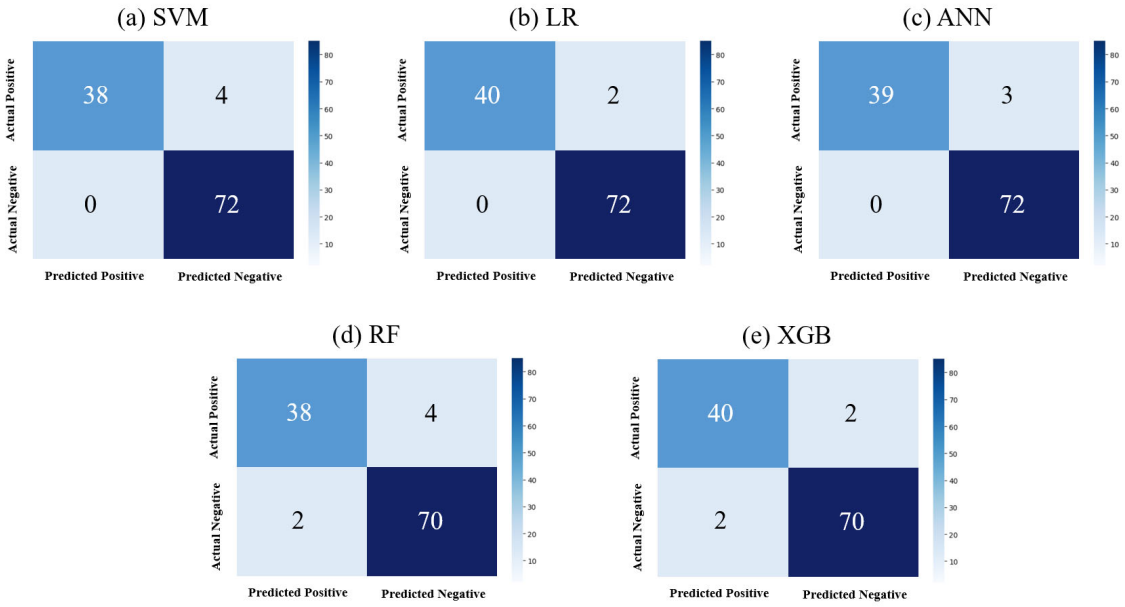

Fig. 5 Confusion matrices of SVM, LR, ANN, RF and XGB

3.2.2 Metrics-based algorithm comparison

Table 2. Comparative evaluation of high PCA accuracy algorithms on key performance metrics

Algorithm	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	AUC
SVM	97.37%*	97.62%*	95.35%*	96.47%*	94.37%*	99.74%
LR	97.37%*	97.62%*	95.35%*	96.47%*	94.37%*	99.77%*
ANN	96.49%	95.35%	95.35%*	95.35%	92.53%	99.51%
RF	96.49%	97.56%	93.02%	95.24%	92.46%	99.28%
XGB	95.61%	95.24%	93.02%	94.12%	90.62%	99.41%

* Means the optimum value of 6 indexes among 5 algorithms

In the pursuit of determining the optimal machine learning algorithm for breast cancer diagnosis, various established metrics were conducted on these algorithms. As seen from Table 2, the following observations about the performance of these algorithms were made:

• Accuracy: All algorithms exhibit high accuracy, with SVM and LR achieving the highest accuracy rate of 97.37%.

• Precision: It is an essential metric, especially in applications where false positives might have significant consequences. SVM and LR demonstrated the most excellent precision at 97.62%.

• Recall: SVM, LR, and ANN showcased robust performance with the highest recall rate of 95.35%.

• F1 Score: This is the harmonic mean of precision and recall. The five algorithms showcase high F1 scores, indicating that they are of high quality on the post-PCA data. SVM and LR exhibited superior performance with an F1 score of 96.47%. The outcome above highlights the algorithms' dependable predicted consistency and well-balanced precision-recall performance.

• Cohen Kappa: This metric evaluates the level of agreement between predicted and observed categorisations while accounting for the possibility of agreement occurring by coincidence. Higher values suggest better agreement. All algorithms display high Cohen Kappa scores, but SVM and LR dominated the metrics with a Cohen Kappa score of 94.37%

• AUC Metric Scrutiny: The AUC is an instrumental metric in measuring the integral performance of a classification algorithm. An AUC proximate to 1 is symbolic of an excellent algorithm. In this

context, all algorithms exhibit AUC values in the upper 0.99 echelon, implying exceptional prediction proficiencies. The highest score is for LR at 99.77%, while the SVM model closely follows with a score of 99.74%.

3.2.3 ROC curve

The evaluation of machine learning algorithms on the dataset included the analysis of Receiver Operating Characteristic (ROC) curves, as shown in Fig. 6. The ROC curve visually represents the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) for every potential cut-off. The idea algorithm would have a ROC curve that traverses the upper-left corner, satisfying 100% sensitivity (no false negatives) and 100% specificity (no false positives). The dashed line in the figure signifies the line of no discrimination when the algorithm's ability to differentiate between the positive and negative classfications is tantamount to a random selection.

In Fig. 6, all algorithms demonstrate a significant level of discriminative power, with LR slightly surpassing the others in terms of performance.

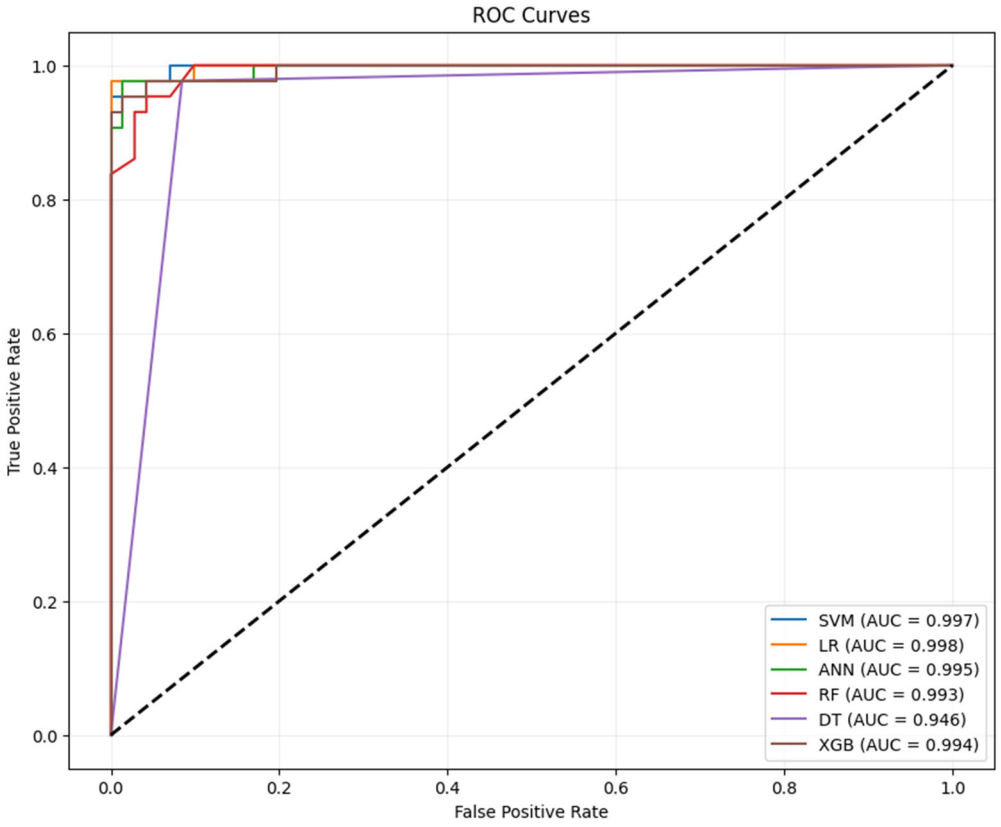

Fig. 6 ROC curves of SVM, LR, ANN, RF and XGB

In summary, all the analysed algorithms demonstrated commendable results on this dataset. However, based on the above metrics, LR and SVM algorithms are the most suitable for implementation. However, the algorithm selection should consider the applications' practical requirements. If computational efficiency is paramount, LR may be preferred, while SVM would be more appropriate in situations where achieving the highest level of accuracy is prioritised. Although ANN and RF demonstrated outstanding results, their measurements were slightly overshadowed by the superior metrics of SVM and LR. In contrast, DT and XGB demonstrate slightly lower metrics despite notable accuracy and other indications, suggesting their potential usefulness in other circumstances or datasets.

4. Discussion

In this study, Genetic Algorithms (GA) and Model Feature Importance were first regarded as techniques for dimensionality reduction. However, it was shown that both approaches had some limitations. GA operates as a global search method, exploring various regions inside the feature space to discover the most optimum combinations of features. Nevertheless, the stochastic nature of GA may result in variable outcomes over several runs, typically necessitating longer calculation durations and too much randomness. In contrast, Model Feature Importance has notable benefits of speed and interpretability, such as the direct provision of feature importance scores by RF. However, the outcomes significantly depend on the algorithm used, resulting in disparate importance scores across various algorithms. Considering the abovementioned obstacles and the successful implementation of PCA in previous works like "Machine Learning Classifiers on Breast Cancer Recurrences" [16]. PCA was chosen as the most suitable method for reducing the dimensionality in this study.

To enhance the accuracy of the selection process among GA, Model Feature Importance, and PCA, it is essential to conduct cross-validation utilising the features selected by each approach. Additional factors, such as model interpretability and time and computational constraints, significantly shape the decision-making process. The multifaceted evaluation guarantees ensure the best technique is chosen to achieve reliable results in the context of breast cancer recurrences.

One notable limitation when assessing the present dataset is its comparatively limited scale, comprising just 569 data points. More data may help the strength and applicability of machine learning algorithms. Although smaller datasets may be suitable for some statistical methods, modern machine learning algorithms, especially deep learning variations, often need larger amounts of data to effectively understand complex patterns and assure the accuracy of the obtained insights. Moreover, the potential for overfitting is exacerbated with limited data. Hence, it is crucial to acquire more comprehensive datasets for training and testing. By engaging in this practice, one may enhance the precision of algorithms and get a comprehensive understanding of the fundamental patterns and connections involved. It would be advisable to explore acquiring larger datasets or ways for further data augmentation to strengthen this study.

5. Conclusion

This study's objective was to comprehensively evaluate the efficacy of seven machine learning algorithms, namely ANN, SVM, DT, RF, XGB, LR, and KNN, in the context of the Breast Cancer Wisconsin (Diagnostic) Database. A comprehensive set of metrics was used for evaluation, including accuracy, precision, recall, F1 score, Cohen's Kappa, and AUC. All algorithms were built using features transformed by PCA. The study results demonstrate that LR exhibited superior performance compared to other algorithms. LR achieved an accuracy rate of 97.37%, precision of 97.62%, recall of 95.35%, F1 score of 96.47%, Cohen's Kappa coefficient of 94.37%, and an AUC of 99.77%. Significantly, the SVM algorithm exhibited a performance that was substantially indistinguishable from that of LR, as shown by an AUC value of 99.74%. This AUC value is slightly lower by 0.03% compared to LR, while all other metrics remained consistent with those of LR. Hence, it is evident that both LR and SVM provide the most appropriate algorithms for prediction, demonstrating remarkably similar high-quality performance measures.

Based on the obtained findings, the study provides evidence supporting the effectiveness of LR and SVM algorithms in detecting breast cancer recurrences. However, it also recognises the importance of selecting the most suitable algorithm based on the specific context, considering practical factors such as computational costs and required levels of accuracy. It is crucial to recognise that the dataset used in the study consisted of only 569 data points, hence imposing limitations on the algorithms' resilience and generalizability. To build upon this study, future endeavours may strive to acquire larger datasets and further examine the effectiveness of different algorithms, particularly when applied to smaller datasets.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394-424.

[2] Breast cancer. 2023, July 12. Breast Cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[3] Rashmi R, Prasad K, Udupa C B K. Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review. Journal of Medical Systems. 2021, December 3; 46(1).

[4] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med. 2019 Jan; 25(1):24-29.

[5] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015; 61: 85-117.

[6] Wolberg William, Mangasarian Olvi, Street Nick, Street W. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. 1995.

[7] Heaton J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. Genetic Programming and Evolvable Machines. 2017, October 29; 19(1–2): 305–307.

[8] Dong S, He D, Zhang Q, Huang C, Hu Z, Zhang C, Nie L, Wang K, Luo W, Yu J, Tian B, Wu W, Chen X, Wang F, Hu J, Xiao X. Early cancer detection by serum biomolecular fingerprinting spectroscopy with machine learning. ELight. 2023, July 24; 3(1).

[9] Noble W S. What is a support vector machine? - Nature Biotechnology. Nature. 2006, December 1.

[10] Quinlan J R. Induction of decision trees. Machine Learning. 1986, March; 1(1): 81–106.

[11] Breiman L. Random forests. Machine Learning. 2001; 45(1): 5-32.

[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016; pp. 785-794.

[13] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression. John Wiley & Sons. 2013.

[14] Altman N S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician. 1992, August; 46(3): 175–185.

[15] Jolliffe I T. Principal Component Analysis, Second Edition. Springer Series in Statistics. New York: Springer. 2002.

[16] Magboo V P C, Magboo M S A. Machine Learning Classifiers on Breast Cancer Recurrences. Procedia Computer Science. 2021; 192: 2742–2752.