International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 2, April 2024, pp. 1937~1947
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i2.pp1937-1947

Predictive modeling for breast cancer based on machine learning algorithms and features selection methods

Arar Al Tawil[1], Laiali Almazaydeh[2,4], Bilal Alqudah[3], Abedallah Zaid Abualkishik[4], Ali A. Alwan[5]

[1]Faculty of Information Technology, Applied Science Private University, Amman, Jordan
[2]College of Information Technology, Al-Hussein Bin Talal University, Ma'an, Jordan
[3]College of Engineering, Al-Hussein Bin Talal University, Ma'an, Jordan
[4]College of Computer Information Technology, The American University in the Emirates, Dubai, United Arab Emirates
[5]School of Theoretical and Applied Science, Ramapo College of New Jersey, New Jersey, United States of America

Article Info

Article history:

Received Sep 17, 2023
Revised Dec 6, 2023
Accepted Dec 18, 2023

Keywords:

Breast cancer
K-nearest neighbors
Light gradient boosting machine
Logistic regression
Naïve Bayes
Random forest
Support vector machine

ABSTRACT

Breast cancer is one of the leading causes of death among women worldwide. However, early prediction of breast cancer plays a crucial role. Therefore, strong needs exist for automatic accurate early prediction of breast cancer. In this paper, machine learning (ML) classifiers combined with features selection methods are used to build an intelligent tool for breast cancer prediction. The Wisconsin diagnostic breast cancer (WDBC) dataset is used to train and test the model. Classification algorithms, including support vector machine (SVM), light gradient boosting machine (LightGBM), random forest (RF), logistic regression (LR), k-nearest neighbors (k-NN), and naïve Bayes, were employed. Performance measures for each of them were obtained, namely: accuracy, precision, recall, F-score, Kappa, Matthews correlation coefficient (MCC), and time. The results indicate that without feature selection, LightGBM achieves the highest accuracy at 95%. With minimum redundancy maximum relevance (mRMR) feature selection (15 features), LightGBM outperforms other classifiers, achieving an accuracy of 98%. For Pearson correlation coefficient feature selection (15 features), LightGBM also excels with a 95% accuracy rate. Lasso feature selection (5 features) produces varied results across classifiers, with logistic regression achieving the highest accuracy at 96%. These findings underscore the importance of feature selection in refining model performance and in improving detection for breast cancer.

Corresponding Author:

Arar Al Tawil
Faculty of Information Technology, Applied Science Private University
Amman, Jordan
Email: ar_altawil@asu.edu.jo

1. INTRODUCTION

The costs of healthcare systems worldwide are increasing due to a rapidly aging population, investments in medical technology developments, and the growing spread of chronic diseases. Chronic diseases cost 3.4 million potential productive life years [1]. Major chronic diseases include chronic respiratory disease, hypertension and diabetes, stroke, cardiovascular disease, and cancer [2], [3].

Cancer is a disease in which some cells in the body cells begin to grow and multiply uncontrollably. A group of rapidly damaged cells may form lump or tumor. Tumor can be of two types: malignant (cancerous) which spread across the body and form new tumors through a process called metastasis or benign (non-cancerous) which stay in one place and cannot invade nearby tissues [4].

Breast cancer is one of the most common causes of women's mortality worldwide. Currently nearly four million women in the United States alone are diagnosed with breast cancer [5]. However, the mortality rate can be reduced if breast cancer is diagnosed at an early stage in screening and treated in time. Many years ago, screening trials have been established to improve early detection of breast cancer, but it results in numerous scans that need to be evaluated which is labor intensive. Therefore, strong needs exist for automatic accurate early detection and prognostic prediction of breast cancer.

This study aims to develop and evaluate an automatic detection model of breast cancer using artificial intelligence (AI). AI is being used to accelerate cancer research and therapy [6]. Algorithms are used by AI systems to simulate a human's cognitive abilities. To construct these systems, a large amount of data sets are required, which allows the AI to recognize the pattern within the data set, and identify the relationship between the data.

Machine learning (ML) is used to derive AI applications. ML provides statistical tools to explore and analyze particular data. In ML there is three different techniques: supervised, unsupervised, and semi-supervised ML [7]. Notably, the various ML techniques conduct differently and have a few instances where they outperform each other. Therefore, this paper presents a comprehensive empirical comparison of ML techniques combined with features selection methods on breast cancer prediction. We apply ML classifiers to four different sets of features: all 30 features, a set of 15 features selected by the minimum redundancy maximum relevance (mRMR) algorithm, a set of 15 features selected by Pearson correlation coefficient, and a set of 5 features selected by Lasso algorithm, and finally the outcomes of the different ML classifiers for each features selection methods are evaluated based on their performance measures.

Unlike previous approaches, our approach advances the field by presenting a holistic comparison of supervised ML techniques and feature selection methods specifically tailored for breast cancer prediction. This research fills a critical gap by not only developing an automatic detection model but also systematically evaluating its performance under different conditions. This comparative evaluation offers valuable insights into the performance of ML classifiers in breast cancer prediction, contributing to the advancement of automated breast cancer detection and aiming to improve accuracy while reducing the burden on healthcare systems. The subsequent sections elaborate on these contributions, demonstrating the depth and significance of our approach.

This paper is organized as follows: section 2 reviews the related works. Section 3 provides an in-depth outline of the problem, detailing the proposed work, including the dataset, data pre-processing, feature selection algorithms, 10-fold cross-validation, and ML classifiers used. Section 4 demonstrates the experimental results and evaluation. Section 5 outlines the conclusion and future directions for further research.

2. RELATED WORKS

Different works have featured about shortcomings and advantages of using ML methods in order to predict breast cancer disease. The works related to this area is reviewed in brief as follows. In study [8] supervised ML methods such as SVMs, KNN, and logistic regression were combined with the principal components analysis (PCA) for dimensionality reduction to identify breast cancer patients. The experiment was conducted on data from the University of California, Irvine (UCI) repository [9]. The highest accuracy of 92.7% was obtained with SVMs for identifying breast cancer patients by the proposed approach.

For the Wisconsin diagnosis breast cancer (WDBC) dataset in UCI machine learning repository, the work in [10] used semi-supervised ML methods such as decision trees, random forest, and k-nearest neighbors to build prediction system of breast cancer occurrence. The calculated accuracy of the system was found to be 96% assigned to random forest. Yue *et al.* [11] used the voting approach to implement naive Bayes, J48 and SVM in the ensemble technique for predictive analysis of breast cancer. According to the authors, the ensemble method acquired 97.13% accuracy rate.

Banu and Thirumalaikolundusubramanian [12] have affirmed naive Bayes techniques on the prediction of breast cancer and addressed a comparison study on bayes belief network (BBN), Tree augmented naïve Bayes (TAN) and boosted augmented naive Bayes (BAN). Based on their findings using gradient boosting, 94.11% was the highest accuracy rate have been obtained for TAN. Hence, the authors suggested that TAN is the best classifier among naïve Bayes techniques for Wisconsin breast cancer dataset (WBCD).

Basunia *et al.* [13] proposed an ensemble method named stacking classifier which combines SVM, KNN, and random forest classification techniques. The predicted results of these combined techniques were provided as input into logistic regression as meta classifier. Stacking classifier was applied on WDBC and achieved 97.2% accuracy rate for breast cancer prediction.

Ghani *et al.* [14] used Coimbra breast cancer dataset from UCI repository. The process of breast cancer prediction initiated by pre-processing, then significant attributes were extracted by using recursive feature elimination (RFE). RFE uses random forest to select five features out of nine. The experimental outcomes showed that the artificial neural network (ANN) best classified the data into healthy and patients with an accuracy of 80%. Based on the related works in this field that have been cited, apparently, further research in this field is needed to improve the performance of the classification systems so that it can predict breast cancer ideally, in addition to determine the essential features that affect the prediction performance.

3. METHOD

This section discusses the methodology as depicted in Figure 1 through five main phases. These five phases are dataset preparation, features selection, 10-fold cross-validation, machine learning classification, and prediction model evaluation phase. These phases are outlined in Figure 1.

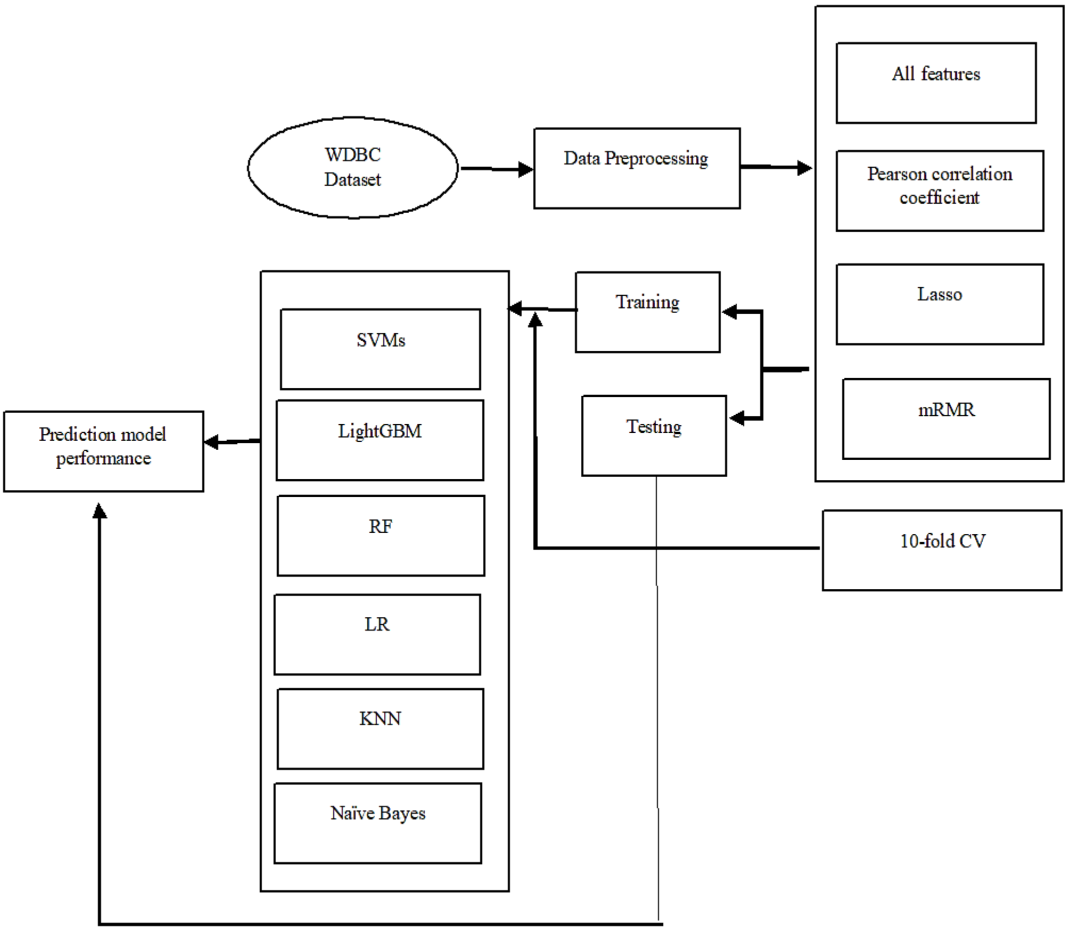

Figure 1. Framework of the proposed methodology

3.1. Dataset preparation

The used dataset of breast cancer is a publicly available dataset called Wisconsin diagnostic breast cancer (WDBC) dataset from UCI machine learning repository in [9]. The dataset contains 569 instances, each instance consists of 30 attributes. Based on these attributes, the diagnosis of the tumor is either Malignant (M) or Benign (B). Table 1 shows the collection of the 30 attributes with their description. In this work, labelling and normalization are the data preprocessing techniques that have been employed for improved ML performance.

3.2. Features selection

Features selection is a process of reducing the input features by selecting only relevant features for ML model training. Using features selection optimizes the training model in several ways: firstly, reducing the dimensionality by removing irrelevant or partially relevant features from the data, hence preventing learning from noise and overfitting, secondly, improving the prediction accuracy, thirdly, reducing the training time, hence it is an exponential growth in some of the training models [15]. There are various statistical methods used for features selection, mainly there are supervised and unsupervised. The former refers to the method for selecting features that makes use of the output label class, Unsupervised features selection, on the other hand, refers to process that does not require the output label class for features selection. Supervised and unsupervised features selection methods can be divided in the same way into three main approaches [16]:

- Filter methods: in these methods, features are eliminated depending on their relation to the output or how they correlate with the output.
- Wrapper methods: in these methods, data are splitted into subsets to train a model, then features are added and deleted based on the model's output to build a subset and train the model again.
- Intrinsic methods: These methods use the benefits of both wrapper and filter and methods to produce the best subset, the model will train and check the accuracy of many subsets before selecting the best one.

Some of the common features selection algorithms based on which method they belong to are [17]:

- Filter methods: Pearson's correlation coefficient, chi squared, ANOVA coefficient.
- Wrapper methods: recursive feature elimination, genetic algorithms.
- Intrinsic methods: lasso regularization, decision tree.

In this paper, based on the input and output variables, as we have a core a numerical input and a numerical output, i.e. not categorical variables, we have used three features selection methods: Pearson's correlation coefficient, lasso and mRMR. These methods are discussed below.

Table 1. WDBC dataset description

Attribute name	Description
Radius mean	Mean of distances from center to points on the perimeter
Texture mean	Standard deviation of gray-scale values
Perimeter mean	Mean size of the core tumor
Area mean	Mean area inside the boundary of core tumor
Smoothness mean	Mean of local variation in radius lengths
Compactness mean	Mean of perimeter^2 / area - 1.0
Concavity mean	Mean of severity of concave portions of the contour
Concave points mean	Mean for number of concave portions of the contour
Symmetry mean	Mean of similar area of tumor parts that matches
Fractal dimension mean	Mean for "coastline approximation" - 1
Radius se	Standard error for the mean of distances from center to points on the perimeter
Texture se	Standard error for standard deviation of gray-scale values
Perimeter se	Standard error for perimeter mean
Area se	Standard error for area mean
Smoothness se	Standard error for local variation in radius lengths
Compactness se	Standard error for perimeter^2 / area - 1.0
Concavity se	Standard error for severity of concave portions of the contour
Concave points se	Standard error for number of concave portions of the contour
Symmetry se	Standard error for mean of similar area of tumor parts that matches
Fractal dimension se	Standard error for "coastline approximation" – 1
Radius worst	"worst" or largest mean value for mean of distances from center to perimeter points
Texture worst	"worst" or largest mean value for standard deviation of gray-scale values
Perimeter worst	"worst" or largest mean value for mean size of the core tumor
Area worst	"worst" or largest mean value for mean area inside the boundary of core tumor
Smoothness worst	"worst" or largest mean value for local variation in radius lengths
Compactness worst	"worst" or largest mean value for perimeter^2 / area - 1.0
Concavity worst	"worst" or largest mean value for severity of concave portions of the contour
Concave points worst	"worst" or largest mean value for number of concave portions of the contour
Symmetry worst	"worst" or largest mean value for standard error for mean of similar area of tumor parts that matches
Fractal dimension worst	"worst" or largest mean value for "coastline approximation" - 1

3.2.1. Pearson correlation coefficient

The Pearson correlation coefficient is often used to assess the strength and direction of a two-variables linear relationship. This coefficient ranges from -1, a perfect negative relationship, to 1, a perfect positive relationship, with 0 representing no correlation [18]. To calculate this coefficient, the formula for the covariance between two variables is used as (1).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

where, r is correlation coefficient, x_i is values of the x-variable in a sample, $\bar{x}$ is mean of the values of the x-variable, y_i is values of the y-variable in a sample, and $\bar{y}$ is mean of the values of the y-variable.

3.2.2. Lasso

Lasso stands for "least absolute shrinkage and selection operator". The concept of lasso regularization is one which is widely utilized in the field of supervised learning. It is a statistical technique which can be used to shrink or regularize the values of coefficients in a given model. By shrinking these coefficients, the model will be better able to reduces the number of variables that need to be considered in the model and reduce the risk of overfitting. Furthermore, it helps to reduce the variance of the model, as the coefficients for insignificant features are shrunk to zero [19].

3.2.3. Minimum redundancy-maximum relevance

It is one of the most heavily utilized features selection algorithms, which uses the concept of mutual information by measuring relevance which is the relationship between each of the features and the class label as well as calculating the redundancy which is a relationship between one variable and another.

First of all, the mRMR run the features in the same manner, and create an empty subset, then looks at the mutual information between all of the features and the class label and try to find which feature has the highest mutual information with the class label to put that feature as the first feature in an empty subset, then more features will be added on the condition that the mutual information between these features that will be added should be minimal with respect to the already selected feature in the subset, and the algorithm continues adding more features sequentially [20]–[22].

3.3. 10-fold cross-validation

10-fold cross-validation is an essential technique used in the field of ML for data analysis and validation. It is a resampling method that divides the dataset into 10 equal parts, or "folds". Each fold is used as a testing set while the remaining nine are used as training sets. After all the folds have been tested, the results are then averaged to get a more reliable estimate. This technique has many advantages, such as reducing the bias of the model and ensuring that the model is not overfitting the data. Additionally, this method allows for the training and testing of the model on the same dataset, which is beneficial in scenarios where the data is limited. Overall, 10-fold cross-validation is a practical and reliable approach for assessing the performance of ML algorithms and its use should be highly considered in a variety of data analysis tasks [23].

3.4. Machine learning algorithms

Machine learning algorithms are especially useful in the field of artificial intelligence. Where they can be used in predictive analytics to analyze large datasets in order to identify patterns, trends, and correlations. The different machine learning classifiers which have been adopted for breast cancer prediction task are discussed below.

3.4.1. Support vector machines

Support vector machines (SVM) is a supervised learning algorithm, it requires a training set or a collection of points that have previously been labeled with the correct category. Each object to be categorized is represented as a point, and the coordinates of a point in an n-dimensional space are frequently referred to as features. SVM performs the classification test by drawing a hyperplane, which is a line in two dimensions or a plane in three dimensions, with all points from one category on one side and all points from the other category on the other. While there may be numerous such hyperplanes, SVM seeks the one that best separates the two categories, and maximizing the distance to points in either category; this distance is known as the margin, and the points that fall exactly on the margin are known as the supporting vectors [24]. In our work, to perform our SVM analysis, we used Python Scikit-Learn. With the final parameters:

params: {C: [10], gamma: 'scale', kernel: ['rbf'], tol: 1e-3}

3.4.2. Light gradient boosting machine

Boosting is an ensemble technique for creating a collection of predictors whose predictions are generally aggregated by some sort of weighted average in order to create an overall prediction that is guided by the collection itself. Gradient boosting is an instantiation of this idea for creating regression models comprised on regressors collections, the idea is repeatedly following this procedure: a simple regression predictor is learned from data, then the error residual is computed, i.e., the amount of the error per data points in the predictions and then a new model to try to predict this error residual is added [25]. In our work, to perform our LightGBM analysis, we used Python Scikit-Learn. With the final parameters:

params: {num_leaves: 31, bagging_freq: 1.0, objective: regression, bagging_fraction: 1.0, learning_rate: 0.1, feature_fraction: 1}

3.4.3. Random forest

Random forest (RF) is a method that works by building several decision trees throughout the training phase. The random forest selects the majority choice of the trees as the final decision. A decision tree is a diagram in the shape of a tree that is used to choose a course of action. Each branch of the tree represents a potential decision occurrence or reaction [26]. In our work, to perform our RF analysis, we used Python Scikit-Learn. With the final parameters:

params: {n_estimators: 100, criterion:'gini', min_samples_leaf:5, min_samples_split:2, max_features:'sqrt'}

3.4.4. Logistic regression

LR derived from the logit transformation which is used in the background of the regression. It is used to describe data and to investigate the connection between one or more independent variables and one or more dependent variables (nominal, ordinal or interval). LR is a technique in which the binary dependent variable can be modeled as a probability of an event-rather than a measure, the measure is always 0 and 1 but the probabilities will always range from 0 to 1, so minimum probability of any event is 0, the max is 1, so this requires the transformation from the binary nominal variable in dataset [27]. In our work, to perform our LR analysis, we used Python Scikit-Learn. With the final parameters:

params: {penalty: 'l2', criterion: 'gini', tol:1e-4, solver:'lbfgs'}

3.4.5. K-nearest neighbors

K-NN is a classification algorithm that stores all existing cases and uses a similarity measure to classify the new data or cases. Therefore, k in k-NN is number of the nearest neighbor which are voting on the data class. The value of this k is a hyperparameter that is chosen by the user precisely such that there is no major bias in one side or the other which resulted in better accuracy [28]. In our work, to perform our k-NN analysis, we used Python Scikit-Learn. With the final parameters:

params: {n_neighbors:5, weights:'uniform', algorithm:'auto'}

3.4.6. Naïve Bayes

Naïve Bayes is a supervised learning algorithm used for classification. It works on the principles of conditional probability as given by the Bayes' theorem. When the dimensionality of the inputs is high, naïve Bayes is an appropriate choice. Despite its simplicity, naïve Bayes may frequently outperform more advanced classification algorithms as it uses fewer training data to predict the classification parameters.

The key assumption of the naïve Bayes algorithm is that class conditional independence, which allows it to simplify the computation of the probabilities. There are several variations of the naïve Bayes classifier that can be used to predict continuous, categorical or binary outcomes. In this work, Gaussian Bayes Naïve is used. In Gaussian Bayes, naïve continuous values associated with each feature are considered to have a Gaussian distribution. A Gaussian distribution is also known as the normal distribution. It generates a symmetric bell-shaped curve around the mean of the feature values when plotted [29]. In our work, to perform our naïve Bayes analysis, we used Python Scikit-Learn. With the final parameters:

params: {var_smoothing=1e-09}

4. RESULTS AND DISCUSSION

In this paper, SVMs, LightGBM, RF, LR, k-NN, naïve Bayes classification algorithms were employed, and classification performance measures for each of them were obtained, namely: accuracy, precision, recall, F-score, MCC, kappa and time. These measures are based on four possible outcomes, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) [30]. The formulae for classification performance metrics are presented in (2)-(7).

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{2}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{4}$$

$$F1\ Score = 2 * \frac{(Recall*Precision)}{(Recall+Precision)} \tag{5}$$

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+TN)}} \tag{6}$$

$$k = \frac{p_o-p_e}{1-p_e} \tag{7}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement.

Each classification algorithm was applied to a set of features selected by three different features selection method. Namely: Pearson correlation coefficient, Lasso, and mRMR, as shown in Table 2. The performance for each of the different classification algorithms and different features selection method is summarized in Table 3 and plotted as shown in Figure 2 to Figure 5.

Based on the experimental results which are tabulated in Table 3, we infer that the lowest accuracy is obtained when all features were used, and specifically with SVMs classification algorithm with an accuracy of 90%. Furthermore, we infer that the highest accuracy is obtained with the combination of the LightGBM classifier and the features selected by mRMR achieving 98%. Comparing these results to different research in the literature demonstrated on WDBC breast cancer, it can be noted that the combination of the LightGBM classifier and the features selected by mRMR outperforms other classifiers in terms of accuracy in breast cancer prediction.

Table 2. Selected features by features selection methods

Algorithms	Features
Pearson correlation coefficient	radius_mean, perimeter_mean, area_mean, compactness_mean, concavity_mean, concavepoints_mean, radius_se, perimeter_se, area_se, radius_worst, perimeter_worst, area_worst, compactness_worst, concavity_worst, concave, points_worst
Lasso	area_mean, area_se, texture_worst, perimeter_worst, area_worst
mRMR	concave points_worst, perimeter_worst, concave points_mean, radius_worst, perimeter_mean, area_worst, radius_mean, concavity_mean, concavity_worst, area_mean, compactness_worst, compactness_mean, texture_worst, radius_se, perimeter_se

Table 3. Selected features by features selection methods

Features selection methods	# Features	Models	Accuracy	Precision	Recall	F-Score	Kappa	MCC	Time
Without features selection	30	Support vector machine	0.90	0.93	0.81	0.85	0.78	0.80	0.01
		Light gradient boosting	0.95	0.94	0.92	0.93	0.89	0.89	0.1
		Random forest classifier	0.96	0.95	0.93	0.94	0.91	0.91	0.5
		Logistic regression	0.95	0.94	0.93	0.93	0.90	0.90	0.20
		K neighbors classifier	0.94	0.94	0.90	0.92	0.87	0.88	0.11
		naïve Bayes	0.93	0.92	0.89	0.90	0.85	0.86	0.01
mRMR	15	Support vector machine	0.93	0.98	0.84	0.90	0.86	0.87	0.01
		Light gradient boosting	0.98	0.98	0.95	0.97	0.95	0.95	0.04
		Random forest classifier	0.95	0.96	0.92	0.94	0.88	0.89	0.01
		Logistic regression	0.95	0.95	0.92	0.94	0.90	0.90	0.2
		K neighbors classifier	0.94	0.97	0.89	0.92	0.88	0.89	0.02
		naïve Bayes	0.94	0.93	0.90	0.91	0.87	0.87	0.01
Pearson correlation coefficient	15	Support vector machine	0.89	0.87	0.86	0.86	0.77	0.78	0.01
		Light gradient boosting	0.95	0.95	0.92	0.93	0.89	0.90	0.1
		Random forest classifier	0.96	0.96	0.93	0.94	0.91	0.91	0.5
		Logistic regression	0.94	0.94	0.92	0.92	0.88	0.89	0.2
		K neighbors classifier	0.93	0.94	0.89	0.91	0.87	0.87	0.12
		naïve Bayes	0.93	0.94	0.89	0.91	0.87	0.87	0.01
Lasso	5	Support vector machine	0.83	0.86	0.77	0.77	0.65	0.70	0.02
		Light gradient boosting	0.94	0.94	0.91	0.92	0.88	0.91	0.4
		Random forest classifier	0.94	0.95	0.91	0.92	0.88	0.90	0.5
		Logistic regression	0.96	0.96	0.93	0.94	0.91	0.91	0.03
		K neighbors classifier	0.93	0.92	0.88	0.90	0.86	0.85	0.11
		naïve Bayes	0.93	0.96	0.84	0.89	0.85	0.85	0.01

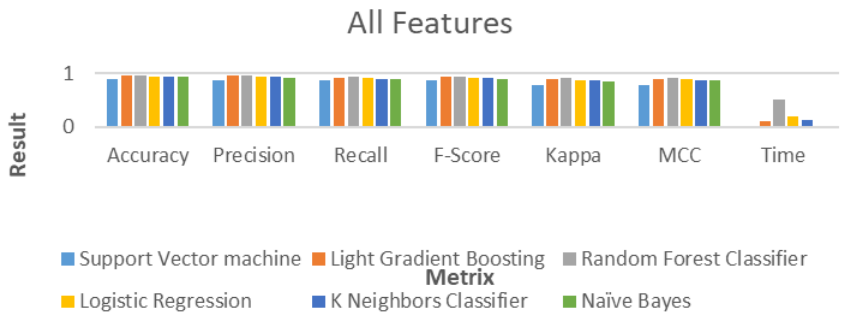

Figure 2. ML algorithms performance with all features

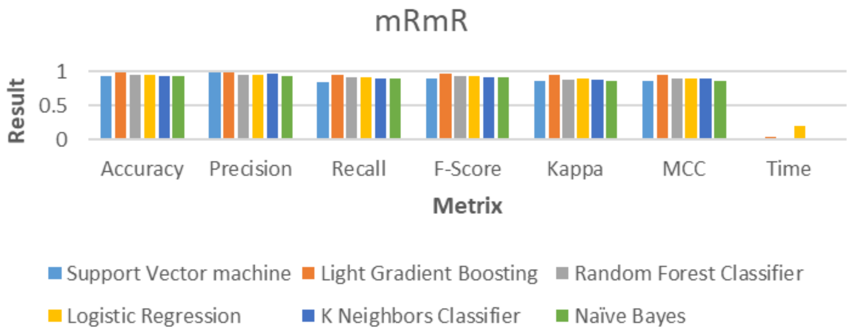

Figure 3. ML algorithms performance with the features selected by mRMR

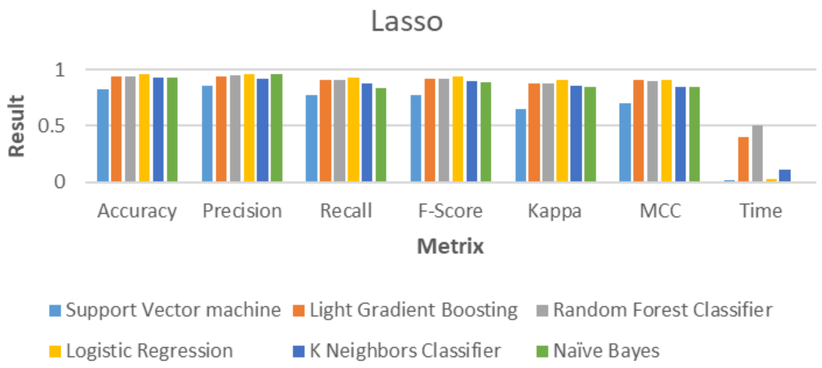

Figure 4. ML algorithms performance with the features selected by Lasso

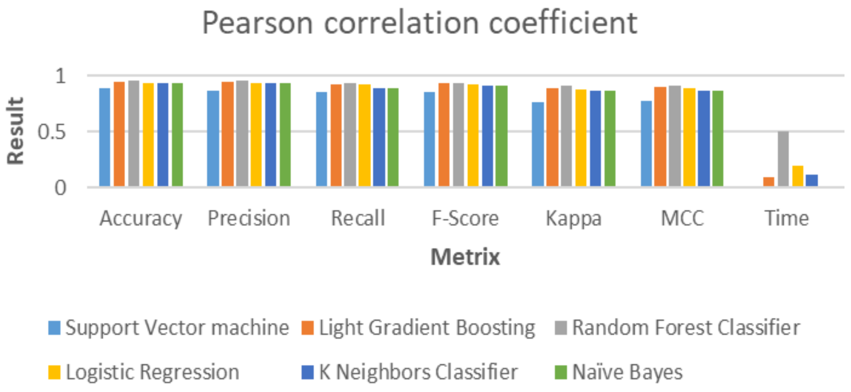

Figure 5. ML algorithms performance with the features selected by Pearson correlation coefficient

5. CONCLUSION AND FUTURE WORKS

In this study, various ML approaches, in conjunction with three common feature selection methods (Pearson correlation coefficient, Lasso, and mRMR), were explored for breast cancer diagnosis using the WDBC dataset. The selected features, crucial in the ML data analysis process, served as inputs for training six ML classifiers: SVMs, LightGBM, RF, LR, k-NN, and naïve Bayes, to create a breast cancer prediction model. The experimental results unveiled compelling insights, showcasing that a model built on features selected by mRMR and the LightGBM classifier achieved the highest accuracy at 98%. This finding signifies the effectiveness of our proposed approach as a valuable tool in the medical field, providing a robust method for breast cancer prediction. Building on the success of the current study, future research avenues can further enhance the proposed approach. Firstly, experiments will be extended to different breast cancer datasets to validate the generalizability and robustness of our methodology across diverse populations and data distributions.

Moreover, future works can delve into the exploration of advanced metaheuristic algorithms for optimizing feature selection and classifier parameters. Metaheuristic algorithms such as particle swarm optimization (PSO), genetic algorithms (GA), and harmony search optimization (HHO) can be employed to fine-tune the selection of features and optimize the performance of ML classifiers. Integrating these metaheuristic algorithms into the feature selection process aims to enhance the efficiency and accuracy of breast cancer prediction models, providing a more sophisticated and adaptive approach. In conclusion, the current research lays the foundation for a promising approach to breast cancer prediction, and future endeavors will focus on extending its applicability, generalizability, and optimization using advanced metaheuristic algorithms.

REFERENCES

[1] Z. Abedjan *et al.*, “Data science in healthcare: benefits, challenges and opportunities,” *Data Science for Healthcare: Methodologies and Applications*, pp. 3–38, 2019, doi: 10.1007/978-3-030-05249-2_1.

[2] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, “Development of disease prediction model based on ensemble learning approach for diabetes and hypertension,” *IEEE Access*, vol. 7, pp. 144777–144789, 2019, doi: 10.1109/ACCESS.2019.2945129.

[3] J. H. Fong, “Disability incidence and functional decline among older adults with major chronic diseases,” *BMC Geriatrics*, vol. 19, no. 1, Nov. 2019, doi: 10.1186/s12877-019-1348-z.

[4] J. Ferlay *et al.*, “Cancer statistics for the year 2020: an overview,” *International Journal of Cancer*, vol. 149, no. 4, pp. 778–789, Apr. 2021, doi: 10.1002/ijc.33588.

[5] D. R. Youlden, S. M. Cramb, N. A. M. Dunn, J. M. Muller, C. M. Pyke, and P. D. Baade, “The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality,” *Cancer Epidemiology*, vol. 36, no. 3, pp. 237–248, Jun. 2012, doi: 10.1016/j.canep.2012.02.007.

[6] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, “AI in health and medicine,” *Nature Medicine*, vol. 28, no. 1, pp. 31–38, Jan. 2022, doi: 10.1038/s41591-021-01614-0.

[7] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, “An ensemble machine learning approach through effective feature extraction to classify fake news,” *Future Generation Computer Systems*, vol. 117, pp. 47–58, Apr. 2021, doi: 10.1016/j.future.2020.11.022.

[8] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, “Prediction of breast cancer using machine learning approaches,” *Journal of Biomedical Physics and Engineering*, vol. 12, no. 3, pp. 297–308, Jul. 2022, doi: 10.31661/jbpe.v0i0.2109-1403.

[9] D. Dua, and C. Graff, “UCI machine learning repository,” Irvine, CA: University of California, School of Information and Computer Science, 2019.

[10] P. Ghosh, M. Z. Hasan, and M. I. Jabiullah, “A comparative study of machine learning approaches on dataset to predicting cancer outcome,” *Journal of the Bangladesh Electronic Society*, vol. 18, no. 1–2, pp. 81–86, 2018.

[11] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, “Machine learning with applications in breast cancer diagnosis and prognosis,” *Designs*, vol. 2, no. 2, pp. 1–17, May 2018, doi: 10.3390/designs2020013.

[12] A. Bazila Banu and P. Thirumalaikolundusubramanian, “Comparison of bayes classifiers for breast cancer classification,” *Asian Pacific Journal of Cancer Prevention*, vol. 19, no. 10, pp. 2917–2920, 2018, doi: 10.22034/APJCP.2018.19.10.2917.

[13] M. R. Basunia *et al.*, “On predicting and analyzing breast cancer using data mining approach,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 1257–1260, doi: 10.1109/TENSYMP50017.2020.9230871.

[14] M. U. Ghani, T. M. Alam, and F. H. Jaskani, “Comparison of classification models for early prediction of breast cancer,” in *2019 International Conference on Innovative Computing (ICIC)*, Nov. 2019, pp. 1–6, doi: 10.1109/ICIC48496.2019.8966691.

[15] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “A review of unsupervised feature selection methods,” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, Jan. 2020, doi: 10.1007/s10462-019-09682-y.

[16] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[17] A. Jovic, K. Brkic, and N. Bogunovic, “A review of feature selection methods with applications,” in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1200–1205, doi: 10.1109/MIPRO.2015.7160458.

[18] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[19] R. Muthukrishnan and R. Rohini, “LASSO: a feature selection technique in predictive modeling for machine learning,” in *2016 IEEE International Conference on Advances in Computer Applications*, Oct. 2017, pp. 18–20, doi: 10.1109/ICACA.2016.7887916.

[20] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and

min redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[21] A. Al-Ani, M. Deriche, and J. Chebil, “A new mutual information based measure for feature selection,” *Intelligent Data Analysis*, vol. 7, no. 1, pp. 43–57, Feb. 2003, doi: 10.3233/ida-2003-7105.

[22] Z. Zhao, R. Anand, and M. Wang, “Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform,” in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2019, pp. 442–452, doi: 10.1109/DSAA.2019.00059.

[23] I. K. Nti, O. Nyarko-Boateng, and J. Aning, “Performance of machine learning algorithms with different K values in K-fold CrossValidation,” *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/ijitcs.2021.06.05.

[24] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[25] C. Chaoura, H. Lazar, and Z. Jarir, “Predictive system of traffic congestion based on machine learning,” in *2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2022, pp. 1–6, doi: 10.1109/WINCOM55661.2022.9966448.

[26] A. Chaudhary, S. Kolhe, and R. Kamal, “An improved random forest classifier for multi-class classification,” *Information Processing in Agriculture*, vol. 3, no. 4, pp. 215–222, Dec. 2016, doi: 10.1016/j.inpa.2016.08.002.

[27] N. Li and R. Jimenez, “A logistic regression classifier for long-term probabilistic prediction of rock burst hazard,” *Natural Hazards*, vol. 90, no. 1, pp. 197–215, Sep. 2018, doi: 10.1007/s11069-017-3044-7.

[28] H. A. Elzeheiry, S. Barakat, and A. Rezk, “Different scales of medical data classification based on machine learning techniques: a comparative study,” *Applied Sciences*, vol. 12, no. 2, Jan. 2022, doi: 10.3390/app12020919.

[29] F.-J. Yang, “An implementation of naive bayes classifier,” in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2018, pp. 301–306, doi: 10.1109/CSCI46756.2018.00065.

[30] L. Almazaydeh, S. Atiewi, A. Al Tawil, and K. Elleithy, “Arabic music genre classification using deep convolutional neural networks (CNNs),” *Computers, Materials and Continua*, vol. 72, no. 3, pp. 5443–5458, 2022, doi: 10.32604/cmc.2022.025526.

BIOGRAPHIES OF AUTHORS

Arar Al Tawil earned his B.Sc. in computer science from Al-Hussein Bin Talal University, Jordan, in 2018, followed by an MSc. from Jordan University in 2021. He currently serves as a lecturer and developer specializing in virtual reality and game design. He holds a prominent position as a lecturer in the esteemed Faculty of Information Technology at Applied Science Private University, Amman. His professional pursuits are deeply rooted in virtual reality, augmented reality environments, and the intricate intersection of machine learning and data analysis. His dedication to these fields is reflected in his ongoing research endeavors, where he continually explores new dimensions of technology. Moreover, he remains at the forefront of innovation and is deeply interested in cutting-edge domains such as deep learning and natural language processing (NLP). This commitment to staying abreast of the latest advancements underscores his dedication to pushing the boundaries of technology and contributing significantly to its ever-evolving landscape. He can be contacted at email: ar_altawil@asu.edu.jo.

Laiali Almazaydeh received her doctorate degree in Computer Science and Engineering from University of Bridgeport in USA in 2013, specializing in human computer interaction. She is currently a full professor and the dean of College of Computer Information Technology, The American University in the Emirates, UAE. Laiali has published more than seventy research papers in various international journals and conferences proceedings, her research interests include human computer interaction, pattern recognition, and computer security. She received best paper awards in 3 conferences, ASEE 2012, ASEE 2013 and ICUMT 2016. Recently she has been awarded two postdoc scholarships from European Union Commission and Jordanian-American Fulbright Commission. She can be contacted at emails: laiali.almazaydeh@ahu.edu.jo, ccit.dean@aue.ae.

Bilal Alqudah received his doctorate degree in computer security and privacy protection from the Bobby B. Lyle College of Engineering, Southern Methodist University in USA in 2015. He is currently an assistant professor of computer security and privacy protection at the college of Engineering at Al-Hussein Bin Talal University, Jordan. Dr. Alqudah has held many local and international training seminars and conferences in his field of specialization. Dr. Alqudah focuses in computer security and privacy research, electronic medical records and access controlling, in addition to other areas of interest. He can be contacted at email: alqu-dah@ahu.edu.jo.

Abedallah Zaid Abualkishik is an accomplished software engineer with a fervent interest in the strategic aspects of software development, coding paradigms. He holds a BSc in Software Engineering from HU, Jordan, and completed both an M.Sc. and Ph.D. in Computer Science with a specialization in software engineering from the esteemed University Putra Malaysia (UPM). A prolific researcher, Dr. Abedallah has authored numerous highly regarded papers published in distinguished, and prestigious international conferences. He actively contributes to the scientific community as a regular reviewer for several prominent journals and conferences. Additionally, he has dedicated his expertise locally, serving as a judge to evaluate projects at various national competitions. Dr. Abedallah is reachable on: Abedallah.abualkishik@aue.ae.

Ali A. Alwan is currently an assistant professor at the School of Theoretical and Applied Science, Ramapo College of New Jersey, United States. He received his master of computer science in 2009 and Ph.D. in computer science in 2013 from Universiti Putra Malaysia (UPM), Malaysia. His research interests include databases (mobile, distributed and parallel), preference queries, web databases, probabilistic, incomplete and uncertain databases, query processing and optimization, data management, data integration, location-based social networks (LBSN), recommendation system, data mining, database in cloud, big data management, and crowd-sourced database. He can be contacted at email: aaljuboo@ramapo.edu.