Journal of Computer Science

Original Research Paper

Breast Cancer Grading using Machine Learning Approach Algorithms

[1]Hiba Nabeel Zalloum, [1]Saada Al Zeer, [1]Amir Manassra, [1]Mutaz Rsmi Abu Sara and [2]Jawad H Alkhateeb

[1]Department of IT, Palestine Ahliya University, Bethlehem, Palestinian Authority
[2]College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Khobar 31952, Saudi Arabia

Article history
Received: 14-07-2022
Revised: 07-10-2022
Accepted: 12-10-2022

Corresponding Author:
Hiba Nabeel Zalloum
Department of IT, Palestine Ahliya University, Bethlehem, Palestinian Authority
Email: hebahzalloum@gmail.com

Abstract: Recently, Breast Cancer (BC) becomes a more common cancer disease in women and it considers the most important sign which leads to death among women. Therefore, it requires efficient methods for detecting it to reduce the risk of death. A positive prognosis and greater chances of survival are improved if the BC is detected early. Currently, machine learning plays an important role in diagnosing BC disease. The various techniques in artificial intelligence and machine learning persuade the researchers in exploring their classification systems in classifying and detecting the BC disease. The algorithms are the K-Nearest Neighbor (KNN), the Support Vector Machine (SVM), random forest, logistic regression, and decision tree. In this study, various algorithms of the machine are proposed in designing the classification system for detecting the BC diseases. To improve the resulting quality, the Principal Component Analysis Algorithm (PCA) is applied. The system was tested and evaluated on the Wisconsin BC dataset from the University of Wisconsin Hospitals. The results were interesting and very good. The accuracy, recall, precision, and F-score of the SVM algorithm were obtained by up to 98% compared to previous work.

Keywords: BC, K-Nearest Neighbor (KNN), Machine Learning, Principal Component Analysis (PCA), Support Vector Machine (SVM)

Introduction

Mainly, cancer disease is one of the greatest threats to human life. BC considers the most common type of cancer disease in women's lives all over the world. The majority of new cancer cases and cancer-related deaths are caused by this disease, making it a public health concern in today's society. A woman's BC is generally regarded as the most common cancer. Mainly, BC has considered the second most cancer-diagnosed disease in women. The early diagnosis of BC can improve the prognosis and chance of survival significantly, early detection helps to recover quickly and continue for longer.

Early detection enhances clinical treatment and makes quick decisions regarding the patient to maintain his health for a longer period (Yue *et al*., 2018). Patients may be saved from unnecessary treatments if tumors are classified accurately. As a result, a substantial study has been done on the proper diagnosis of BC and the classification of patients into benign or malignant categories. Machine Learning (ML) is widely acknowledged as the approach of choice for BC pattern classification and forecast modeling due to its distinct advantages in essential feature detection from complicated BC datasets (Yue *et al*., 2018).

There are many ways to classify information, including classification, machine learning, and artificial intelligence, especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions (Houssami *et al*., 2019). Studies investigating the efficacy of machine learning techniques for BC screening suggest that the technology may one day be as accurate as highly trained radiologists. However, these studies frequently use "enriched" datasets, where the prevalence of cancer is much greater than in groups undergoing screening (up to 55%, as opposed to real-world screening populations, where the prevalence of BC is less than 1%) (Coccia, 2020). In this study, The Machine learning algorithm for classification is applied. In this study, a novel method is classifying the BC disease: Which is based on SVM and the KNN classifiers where the PCA is applied for extracting the features.

Numerous models have been developed in prior work that use various feature sets and machine-learning techniques to diagnose breast cancer (Amrane *et al*., 2018). Both Naive Bayes NB and KNN algorithms were used on the BC Dataset (BCD). A comparison with cross-validation has been suggested. The results showed that KNN gave a higher accuracy with the lowest error rate, as the KNN algorithm gave 97.51 and NB 96.19% (Amrane *et al*., 2018; Ara *et al*., 2021). Analyzed a data set and assessed how well several machine learning algorithms performed in predicting BC from benign or malignant tumors.

The following algorithms were used: SVM, Logistic Regression, KNN, Decision Tree, Naive Bayes classifiers, and Random Forest. With an accuracy of 96.5%, the Random Forest and the Supporting Vector Machine exceeded the. The WBCD dataset was created using data that was gathered from the University of Wisconsin Hospital in Madison, Wisconsin, USA (Ara *et al*., 2021).

İlkuçar *et al*. (2014) used the UCI BC Dataset. There are two types of algorithms for Artificial Neural Networks. Harmony Search and Back Propagation algorithms were used to train the Artificial Neural Networks to feed-forward (ANN). Classification performance was tested utilizing precision, SSE, and regression parameters. The performance values of backpropagation were obtained as 94.1/0.007/0.92 and Harmony Search 97.57/0.005/0.96 respectively (İlkuçar *et al*., 2014).

Douangnoulack and Boonjing (2018) they used PCA to reduce the Wisconsin BC (WBC) dataset of a lossless data reduction technique with good classification performance, to aim at finding the best performance classier by giving minimal classification rules by employing PCA. The J48 decision tree classifier is found to be the best accuracy among the three classifiers: J48 Decision tree 97.36%, Minimized Error Pruning Tree 96.77% and Random Tree 94.72% (Douangnoulack and Boonjing, 2018).

Bayrak *et al*. (2019) used the most popular ML techniques, SVM and ANN in the Wisconsin BC Dataset. The comparison was done on the classification performance of these techniques to each other. By using precision, accuracy, recall, and the ROC area. It was found that the SVM classifier had the best percentage split accuracy of about 95% and the ANN of about 88% (Bayrak *et al*., 2019).

Yedjou *et al*. (2021) presented a novel computer-aided diagnosis system for the prediction, diagnosis, and classification of BC applying ML. In particular, they discussed the concepts of ML and outlined their application in the classification of BC. Using ML approaches, their findings revealed that among the 569 patients involved in this study, 63% were diagnosed with benign tumors and 37% were diagnosed with malignant tumors. Radius, texture, perimeter, area, compactness, concavity, and concave points of the cell are some of the feature characteristics (Yedjou *et al*., 2021).

Materials and Methods

In this study, supervised ML algorithms are used. For training, a labeled dataset was utilized. Technically, 70% of the main dataset is used for training and the rest 30% is used for testing. Combining kernel-based PCA techniques with supervised learning algorithms for classification. Normally this type of technique is used to reduce the dimensionality of a dataset.

The work is conducted on a computer with an HP Core i7, 10 generations, and 8 GB of RAM in the Anaconda environment, which is based on the Python programming language. The BC Wisconsin Diagnostic dataset was used to apply ML classifiers, Decision Tree Random Forests, SVM, Logistic Regression, and KNN. Results were assessed to determine which model has the highest level of accuracy.

Dataset

An innovative and reliable approach for the detection of BC must be created and the dataset is a crucial component of that process. Due to the scarcity of samples and the confidentiality of the patient's data, gathering a dataset is extremely challenging. To create a new and innovative product, the dataset used in this study came from the Wisconsin BC dataset created by Dr. William H. Wolber at the University of Wisconsin Hospitals in Madison. This dataset has 699 data points, including 458 benign tumors, 241 malignant tumors, and 9 characteristics. 16 missing data points are present.

Figure 1 illustrates that 458 (65.5%) tumors are benign tumors and 241 (34.5%) tumors are malign. Meanwhile, Fig. 2 summarizes the process of the model. Finally, the confusion matrix is summarized in Fig. 3.

Missing Values

The researchers handle missing values by summing all numbers of the column and they are divided according to their numbers.

Machine Learning Algorithms

PCA

The PCA reduces the size of the observation space in which given objects are observed. Reduction is achieved by combining new linear variables that describe the items being researched.

DOI: 10.3844/jcssp.2022.1213.1218

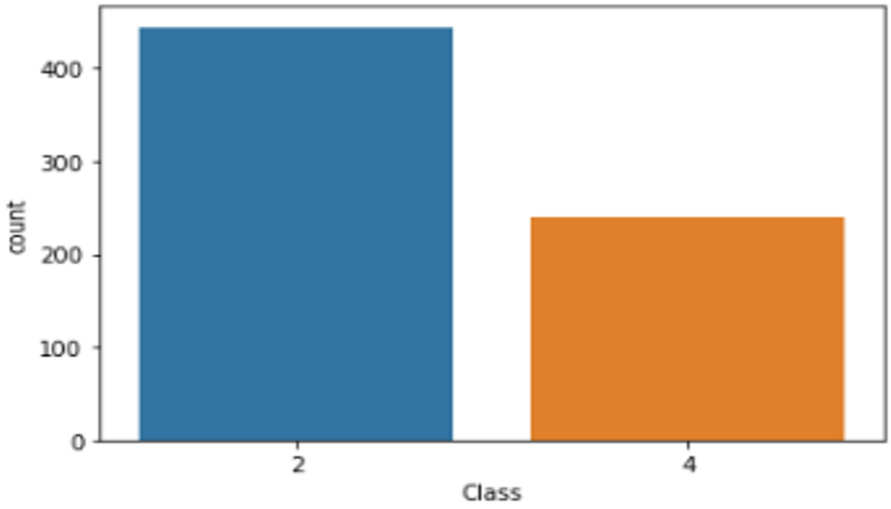

Fig. 1: Wisconsin BC dataset

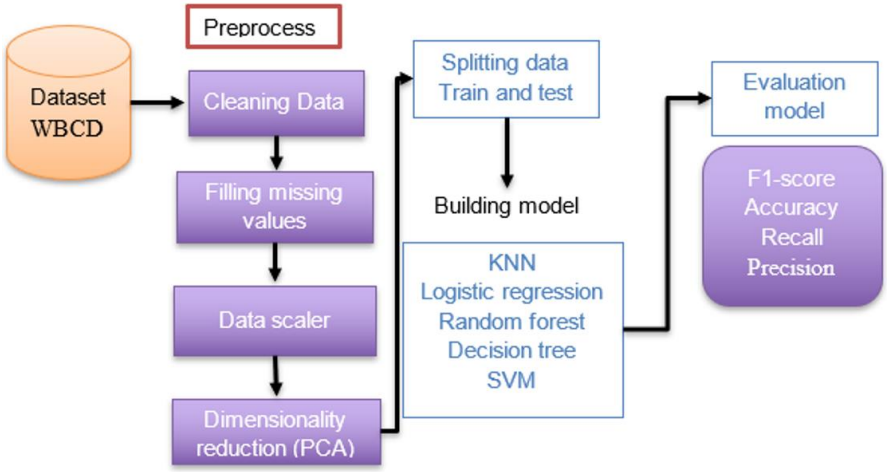

Fig. 2: The process of model

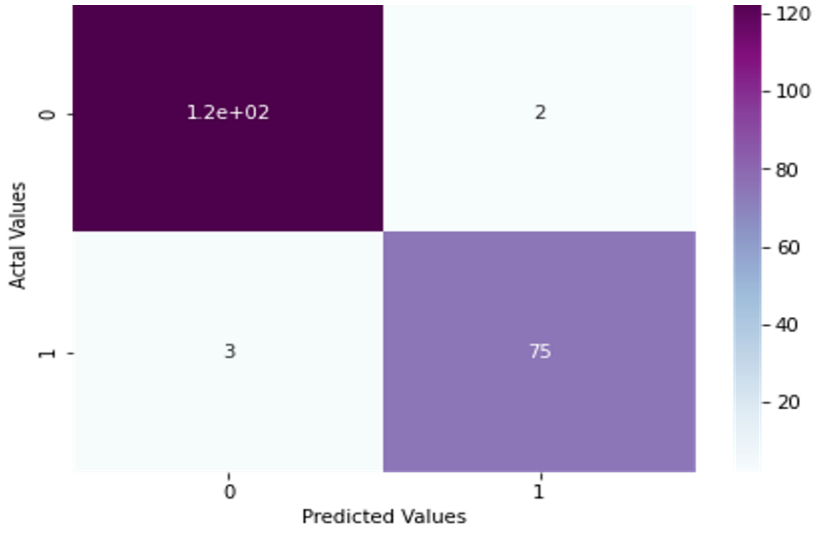

Fig. 3: Confusion Matrix SVM

Combinations that satisfy specific mathematical and statistical requirements are considered principal components (Huang *et al*., 2019; Maćkiewicz and Ratajczak, 1993). The PCA is chosen to remove the weak features and keep the strong ones to get better measurements in terms of accuracy. In essence, the PCA's primary objective is to decrease the amount of data for each dimension. PCA helps the researchers in resizing any image. PCA helps researchers to find various patterns in high-dimensional datasets.

KNN

A highly accurate forecast is what allows an algorithm to compete with the most precise models. In applications that need great accuracy but don't call for a model that can be read by humans, you can therefore utilize the KNN method. The distance metric affects how well the forecasts turn out (WHO, 2020). The equation of Euclidean distance:

$$(x, x_i) = sqrt(sum((x_j - x_{ji})^2 \quad (1)$$

SVM

SVM is one of the most well-liked algorithms for supervised learning and it may be applied to both classification and regression issues. However, it is largely employed in ML Classification issues. The equation of SVM is:

$$B_0 + (B_1 \,.\, X_1) + (B_2 \,.\, B_2) = 0 \quad (2)$$

B_1 and B_2 define the slop of the line and b0 intersection found by the learning algorithm. X_1 and X_2 are the two input variables SVM strike a compromise between two complementary goals, improving accuracy and guaranteeing the highest level of reproducibility. SVM, one of the most well-liked classification techniques in machine learning, has shown beneficial in the detection and prediction of cancer. Using the characteristics of a tumor, this technique has been used to categorize the tumor as benign or malignant (Khourdifi and Bahaj, 2018).

Random Forest

A popular supervised ML algorithm for Classification and Regression issues is Random Forest. On various samples, it constructs decision trees and uses their average for classification and majority vote for regression.

One of the most widely used Machine Learning (ML) methods, supervised learning includes the method of logistic regression. The categorical dependent variable is predicted using a set of independent variables. The equation of logistic regression is:

$$P(x) = e^{\wedge}(b_0 + b_1 \,.\, X) / (1 + e^{\wedge}(b_0 + b_1 \,.\, X)) \quad (3)$$

Decision Tree

A decision support tool known as a decision tree employs a tree-like model to represent options and their potential outcomes, including utility, resource costs, and chance event outcomes. One method for displaying an algorithm that solely uses conditional control statements is to use this one.

Results and Discussion

In this study, Python code was developed to compare the accuracy of several algorithms in the research field to reach the best algorithm, which results in the best accuracy for detecting BC. The samples that were taken for f Patients totaled (699) patients, with (12) important examinations conducted for them (Anji Reddy and Soni, 2020).

After applying ML algorithms to the BC, the Confusion matrix, accuracy, precision, F1 Score, and PCA analysis algorithm are applied to reduce dimensionality. This allows us to choose the optimum algorithm for the BC Prediction by reducing the dimension from 12 to 9 features:

$$Accuaracy = \frac{correct\ predications}{all\ predictions}$$

$$predictions = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ positives}$$

$$F1 - Score = 2 \times = \frac{Precision \times Recall}{Precision + Recall}$$

The effectiveness of a model was evaluated taking into account both precision and recall. Unfortunately, recall and precision can occasionally conflict. To put it another way, improving precision frequently results in a decline in memory and vice versa.

By calculating the harmonic mean of a classifier's precision and recall, the F1-score integrates both into a single metric. It is mainly used to compare the effectiveness of two classifiers. Assume classifiers A and B have higher recall and precision, respectively. The F1-scores for both classifiers in this situation can be used to assess which one yields superior results (Huang *et al*., 2020). Table 1 summarizes the evaluation model results, while Table 2 summarizes the comparison results.

Table 1: Evaluation model result

Model	Accuracy	Precision	Recall	F1-score
LR	0.97	0.97	0.98	0.98
KNN	0.96	0.95	0.98	0.96
SVM	0.98	0.98	0.98	0.98
DT	0.96	0.95	0.98	0.97
RF	0.96	0.95	0.98	0.97

Table 2: Comparing the result

References	Dataset	Algorithms	Accuracy %
Amrane *et al*. (2018)	WBCD	KNN NB	97.51
			96.19
Ara *et al*. (2021)	WBCD	LR KNN DT NB RF SVM	94.40
			95.80
			95.90
			92.30
			96.50
			96.50
Douangnoulack and Boonjing (2018)	WBCD	J48 DT RF	97.36
			94.72
Bayrak *et al*. (2019)	WBCD	SVM ANN	95.00
			88.00
Our research	WBCD	LR KNN SVM DT RF	97.00
			96.00
			98.00
			96.00
			96.00

Comparing the Results with Previous Research

Mainly, what we have discovered in this study reveals that the PCA method, along with the techniques of Logistic Regression, Support Vector Machine, and Random Forest, helped us achieve very accurate findings. The best result we obtained was accuracy from SVM 98%. We obtained a higher accuracy than previous research and this is due to the reason for reducing the size of the data and removing the weak features from it and keeping the strong ones.

Conclusion

An analysis of BC Diagnostic is used in this essay. To identify the best machine learning algorithm that is accurate, dependable, and finds higher accuracy, five main algorithms were primarily applied: SVM, Random Forests, Logistic Regression, Decision Tree, and K-NN. These algorithms calculated, compared, and evaluated various results obtained based on confusion matrix, accuracy, sensitivity, and precision. The sklearn package was used in the Anaconda environment to program all algorithms in Python.

We found that SVM had a higher efficiency of 98%, precision of 98%, recall of 98%, and F1-score of 98% after accurately comparing our models. SVM, which offers the best performance in terms of accuracy and precision, has proven to be effective in BC prediction, recall, f1-score, and diagnosis. It should be mentioned that the WBCD database is the exclusive source of all the results (Huang *et al*., 2019; Maćkiewicz and Ratajczak, 1993).

The rate of improvement in prediction and the accuracy of the result was also increased by 0.5%, which is a high and very important percentage in predicting the presence of BC or not, as the highest percentage was previously 0.975% and after that improvement, we made 98%, to eliminate as many patients as possible by obtaining the highest accuracy in predicting the outcome.

Acknowledgment

We are extremely grateful to "Palestine Al-Ahliya University" and the study's principal investigator, Dr. Mutaz Abu Sara, for their steadfast assistance and extended support in distributing this research report. Additionally, Dr. William H. Wolberg wishes to extend his gratitude to the University of Wisconsin Hospitals Madison for providing the dataset.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Hiba Nabeel Zalloum: Suggested add PCA the dimensionality reduction algorithm and apply it to the machine learning algorithms. She studied previous research and suggested a new contribution.

Saada Al Zeer and Amir Manassra: Conception a design of the article and studied previous research, working on graphs, delving into the results of algorithms, analyzing them, and extracting the final results, equations, and tables of results.

Mutaz Rasmi Abu Sara and Jawad H Alkhateeb: They supervised the work done in this study and drafted the paper in terms of technical and English proofreading for important intellectual content.

Ethics

The corresponding author confirms and attests that other authors reviewed and approved the work and that there were no ethical dilemmas. The reference section included complete and accurate citations for each source.

References

Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE. https://doi.org/10.1109/EBBT.2018.8391453

Anji Reddy, V., & Soni, B. (2020). Breast Cancer Identification and Diagnosis Techniques. In *Machine Learning for Intelligent Decision Science* (pp. 49-70). Springer, Singapore. https://doi.org/10.1007/978-981-15-3689-2_3

Ara, S., Das, A., & Dey, A. (2021, April). Malignant and benign breast cancer classification using machine learning algorithms. In *2021 International Conference on Artificial Intelligence (ICAI)* (pp. 97-101). IEEE. https://doi.org/10.1109/ICAI52203.2021.9445249

Coccia, M. (2020). Deep learning techniques for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. *Technology in Society*, *60*, 101198. https://doi.org/10.1016/j.techsoc.2019.101198

Douangnoulack, P., & Boonjing, V. (2018). Building minimal classification rules for breast cancer diagnosis. In *2018 10th International Conference on Knowledge and Smart Technology (KST)* (pp. 278-281). IEEE. https://doi.org/10.1109/KST.2018.8426198

Houssami, N., Kirkpatrick-Jones, G., Noguchi, N., & Lee, C. I. (2019). Artificial Intelligence (AI) for the early detection of breast cancer: A scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices*, *16*(5), 351-362. https://doi.org/10.1080/17434440.2019.1610387

Huang, Q., Chen, Y., Liu, L., Tao, D., & Li, X. (2019). On combining biclustering mining and AdaBoost for breast tumor classification. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 728-738. https://doi.org/10.1109/TKDE.2019.2891622

İlkuçar, M., Işik, A. H., & Çifci, A. (2014, April). Classification of breast cancer data with harmony search and backpropagation based artificial neural network. In *2014 22nd signal processing and Communications Applications Conference (SIU)* (pp. 762-765). IEEE. https://doi.org/10.1109/SIU.2014.6830341

Bayrak, E. A., Kırcı, P., & Ensari, T. (2019). Comparison of machine learning methods for BC diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). IEEE. https://ieeexplore.ieee.org/document/8741990

Khourdifi, Y., & Bahaj, M. (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (pp. 1-5). IEEE.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R

WHO. (2020). Preventing. World Health Organization. cancer. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en

Yedjou, C. G., Tchounwou, S. S., Aló, R. A., Elhag, R., Mochona, B., & Latinwo, L. (2021). Application of machine learning algorithms in breast cancer diagnosis and classification. *International Journal of Science Academic Research*, *2*(1), 3081.

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13. https://doi.org/10.3390/designs2020013