A PROPOSED TECHNIQUE FOR BREAST CANCER PREDICTION AND CLASSIFICATION BASED ON MACHINE LEARNING

Raneem Ahmed Hegazii[1], Eman Abdelhalim[2], Hossam El-Din Mostafa[3]

Article History: Received: 10.05.2023 **Revised:** 25.06.2023 **Accepted:** 01.07.2023

Abstract
One of the most prevalent forms of cancer among women is breast cancer. Early and precise detection can minimize the impact on the health of patients. Therefore, Machine learning approaches can substantially improve the process of early cancer diagnosis and prediction. This study focuses on the use of machine learning techniques for the prediction and detection of breast cancer. The proposed model involves applying
a set of nine distinct ML based classification models such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost) and Artificial Neural Network (ANN). Several experiments were conducted in this study including different data splitting sizes and feature selection methods. The assessment of these models has been done based on four performance metrics including accuracy, precision, f-measure, and recall. Results indicated that KNN yielded the highest scores of 97.37% and 100% in terms of accuracy and precision respectively in less computational time. Logistic Regression achieved 98.24% in terms of accuracy while using a fewer number of predictive features. While the neural network reached the highest accuracy of 98.25% and outperformed the remaining techniques.

Keywords: *Breast Cancer, Machine Learning, Feature Selection, Performance Measures, Classification.*

1 Electronics and Communication Engineering Department, Faculty of Engineering, Horus University, Egypt,
2 Electronics and Communication Engineering Department, Faculty of Engineering, Mansoura University, Egypt
3 Electronics and Communication Engineering Department, Faculty of Engineering, Mansoura University, Egypt
Corresponding Email : *raneemhegazy5@gmail.com*

DOI: 10.48047/ecb/2023.12.8.619

1. Introduction

Cancer is a major public health problem around the world. Breast cancer is the most common and leading cause of cancer among women, and it is still rising in both the developing and developed worlds. In 2012, it represented about 12 percent of all new cancer cases and 25 percent of all cancers in women Breast cancer is one of the most lethal and heterogeneous diseases in this present era that causes the death of an enormous number of women all over the world. It is the second largest disease that is responsible for women's death. In 2020, more than 2.3 million women were diagnosed with breast cancer worldwide and 685,000 died. Every 14 seconds, somewhere in the world, a woman is diagnosed with breast cancer. Globally, breast cancer now represents one in four of all cancers in women. Since 2008, worldwide breast cancer incidence has increased by more than 20 percent. Mortality has increased by 14 percent. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer it is caused by the rapid and unstable division of breast cells, which results in a lump in the breast. It then travels to other parts of the body via the lymph nodes. This condition can affect the milk ducts, glandular tissue, or other breast tissue. All women are advised to check for signs of breast cancer on a regular basis and to consult their doctor if there is a lump in the breast or if there is any change in the tissue. Early detection of breast cancer can increase recovery rates worldwide. The rapid progress of computer science and algorithms has enabled novel approaches to harnessing data in order to discover more insight for competitive advantages. For the prediction of breast cancer, various machine learning and data mining algorithms are being used. One of the most important tasks is to find the best and most appropriate algorithm for predicting breast cancer. Machine learning is one of the most rapidly growing fields of computer science. Its main goal is to enable computers to learn from input data, commonly referred to as training data, and extract knowledge to perform tasks on future data. Learning is classified into three types: supervised, unsupervised, and reinforcement learning. There are several techniques and algorithms for each type.

Machine learning techniques had been used to improve diagnosis speed and accuracy.

Abdar et al. [1] achieved an accuracy of 98.07% using both Support vector and Naïve bayes techniques, However Naïve bayes achieved it in a less building model time. Nguyen et al. [2] utilized the models of LR, SVM and AdaBoost that scored the highest accuracy of 98%. Chaurasia et al. [3] performed analysis on the original dataset with feature selection using SVM, KNN and Naive Bayes achieving an accuracy of 98.158%, 98.16% and 98.157%, respectively. Basunia et al. [4] suggested an ensemble method to combine the results of the different classifiers and provided 97.20% accuracy for breast cancer detection. Naji et al. [5] utilized five different machine learning models including SVM, RF, LR, DT, and KNN. SVM outperformed all other classifiers, achieving the highest accuracy 97.2%. Khan et al. [6] analyzed dataset and utilized variety of machinelearning models such as RF, LR, DT, and KNN for prediction of breast cancer. When the results were compared, it concluded that LR achieved the best results of 98% accuracy rate. Singh et al. [7] describes a hybrid technique for detecting most important features combined from two different models and achieved astounding results with up to 98.16% accuracy. Bataineh [8] proposed a model that include various machine learning techniques. MLP, KNN, SVM, and NB were compared. MLP achieved an accuracy of 96.70%, which is higher than the other algorithms. Mangukiya et al. [9] performed data visualization and performance comparisons between seven different machine learning models such as SVM, DT, NB, KNN, Adaboost, XGBoost, and RF. Based on the results of the experiments, XGBoost has the highest accuracy 98.14. Table 1 presents a summary of related work.

This study aims to improve the accuracy of prediction model for breast cancer. The proposed methodology is covered in detail in Section 2. Section 3 delves into the evaluation and discussion of experimental results. Work conclusions are discussed in Section 4.

The contributions of proposed work are

1) The use of various machine learning models for classification of breast cancer.
2) Feature selection techniques for choosing the most relevant feature for target classification and excluding the irrelevant or redundant ones.

Table 1. Summary of Related Work

Author	Technique	Accuracy
Abdar et al.	SVM and NB	98.07%
Nguyen et al.	Adaboost	98%
Chaurasia et al.	SVM KNN NB	98.158% 98.16% 98.157%
Basunia et al.	Ensemble	97.2%
Naji et al.	SVM	97.2%
Khan et al.	LR	98%
Singh et al.	Hybrid	98.16%
Bataineh	MLP	96.7%
Mangukiya et al.	XGBoost	98.14%

2. Methods

This section consists of four main blocks of the proposed framework for breast cancer classification as shown in Figure 1. Firstly, it contains a brief description of the datasets. Then it delves into data preprocessing and machine learning models that were implemented in this study. Finally, the evaluation parameters used to assess the performance of the prediction model are discussed.

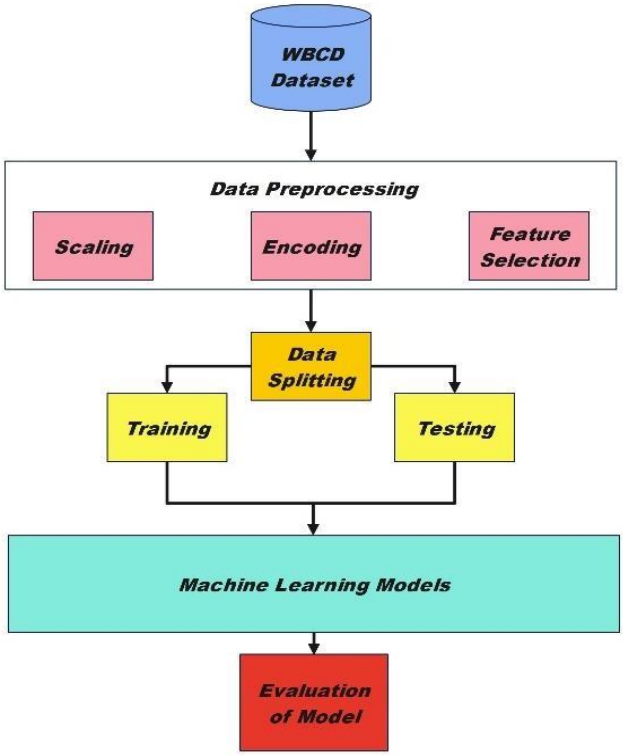

Figure 1. Block diagram of the proposed framework

A. Dataset Description

The dataset implemented in this research is available online on UCI machine learning respiratory named Wisconsin Breast Cancer Dataset (WBCD). It consists of 569 instances with 32 attributes. Dataset includes 357 cases that were identified as benign and 212 were classified as malignant. Distribution of the dataset is shown in Figure 2.

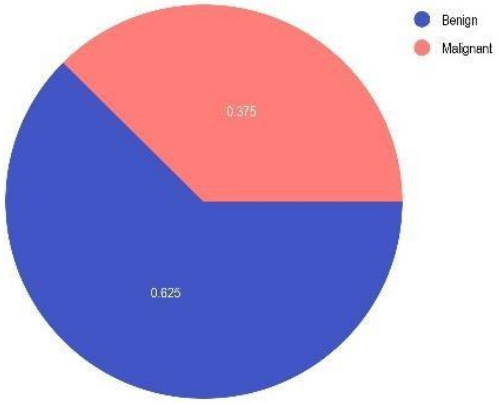

Figure 2. Dataset Distribution

B. Data preprocessing

Preprocessing is an important building block in the prediction model. Data preprocessing is implemented in three different stages which are encoding, scaling and feature selection.

- *Encoding:* Nominal values need to be converted into numbers to make machine learning algorithm able to understand data it receives in order to facilitate processing. Categorical variables in this research were encoded by a label encoder technique. Where each value is assigned a unique integer.
- *Scaling*: Features with larger values or ranges may dominate the learning process and lead to a biased model. Therefore, each data point is resized in a certain range during normalization using the following equation:

$$y = \frac{x - Min.Value}{Max\ value - Min.\ Value} \quad (1)$$

Where; x is the value before normalization, and y is the value after normalization.

- *Feature Selection:* It is a technique used for reducing the number of not-related input variables that do not have a powerful contribution on the target classification variable. It focuses on the problem of high complexity to yield better relative accuracy[10]. Therefore, the model runs more quickly and the dimensionality of the data is reduced. Feature selection was implemented in this research using the ANOVA for continuous numerical features

C. Machine Learning Models

Machine learning is a subfield of artificial intelligence which is broadly defined as a machine's capacity to reproduce intelligent human behavior. Machine learning constructs a model from sample data, referred to it as training data, in order to make predictions or decisions without being explicitly programmed to do so. Supervised learning and unsupervised learning are the two types of machine learning. Since dataset used in this study contains labeled data, nine supervised machine learning techniques were implemented. Methodology includes Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN) and Neural Network (NN).

1. **Naive Bayes**: A supervised technique that requires feature independence for data classification is a probability-based model. For datasets with a large number of input attributes, this model performs well [12]. Every available feature is included, even those that very slightly affect the outcome of the prediction.
2. **Support Vector Machine:** One of the most reliable statistical learning frameworkbased algorithms. It is regarded as a Decision plane-based model [13] that offers a solution for both regression and classification problems as well as for both linear and non-linear datasets.
3. **Logistic Regression:** A well-known supervised learning approach in the medical field. The probability of the class output is predicted using logistic regression utilizing a collection of independent characteristics [14]. If p is the likelihood that a subject falls into the benign class, then 1-p represents the likelihood that a subject falls into the malignant class. The threshold used to decide which data belongs to a particular class is known as a decision boundary.
4. **Decision Tree:** A Hierarchy-based model that aims to understand fundamental chained decision rules from previous input variables in order to train a model to categorize a target variable. The variables are recursively separated using a set of impurity criteria up until a set of stopping conditions are satisfied [15].
5. **Random Forest:** During the training phase, RF creates a large number of decision trees and then produces classes for each of those trees. Both classification and regression can be used. The de-correlated tree via bagging, which is the creation of multiple decision trees from training data using bootstrapped samples with a minor adjustment [16].
6. **K-nearest neighbors:** A supervised classification algorithm used in order to recognize patterns. It uses a large number of identified points to teach itself how to label new ones [17]. In order to label a new point, it considers the identified points that are nearby, or its nearest neighbors, and asks those neighbors to make decisions.
7. **Neural Network:** Based on their layers and neurons, NNs would be created for classification, recognition, as well as a variety of other functions. Biological neural networks underlie its fundamental methodology. After being correctly trained, an NN may learn how to perform its functions and generalize to give an acceptable response to unseen data, which are its two important qualities. The best subset of features is initially provided to the NN as inputs. Each neuron computes a weighted sum for each feature subset [18]. This weighted sum is then subjected to a transfer function to ascertain the output value of the neuron. Parameters used in building neural network model includes:

- Batch size: refer to number of samples used in each iteration. Larger batch size requires more memory and computational cost. 16,32, 64.
- Learning rate: step size used to update weights of neural network. Higher learning rate can exceed optimal weights. 0.0001,0.001,0.01,0.1 and 1
- Activation function: mathematical function applied to output of each neuron. ReLU, sigmoid, tanh, and softmax.
- Optimizer: algorithm used to update weights of neural network. Adam, Stochastic gradient descent SGD and Adagrad.
- Epochs: an iteration of training over the entire dataset. Higher number may lead to better performance but overfitting.

8. **Extreme Gradient Boosting:** It is a tree-based sequential decision trees algorithms [19]. It is considered as one of the most efficient methods for performing classification and predictions on structured or tabular datasets. Scalability is regarded as the most significant aspect that it enables direct abrupt learning through parallel computation in addition to providing an optimized memory usage [20].

Adaptive Boosting: It is an iterative machine learning algorithm that is less affected by the problem of overfitting. Where dataset is split into two parts for each iteration, the features used in the first iteration will be given less weight, and the mistakenly classified data are given more weight in

the next round. When all iterations are finally completed, they are merged with appropriate weights to produce an effective classifier that can unseen data classes.

D. Evaluation parameters

- **Confusion matrix** is a certain table structure that allows visualization of the performance of an algorithm, usually one that uses supervised learning cancer diagnosis may be benign or malignant so, the confusion matrix used to define the performance of a classification algorithm. It indicates the number of false positives FP, false negatives FN, true positives TP, and true negatives TN.
- **Sensitivity**:
relates to the test's ability to correctly detect ill patients who do have the condition

$$Sensitivity = \frac{TP}{TN+FP} \quad (2)$$

- **Specificity**:
refers to the test's ability to correctly reject healthy patients without a condition

$$Specificty = \frac{TN}{TN+FP} \quad (3)$$

- **F1 Score:**
It is the weighted average of precision and recall.

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

- **Recall:**
It is the proportion of correctly predicted events among the foreseen data.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

- **Accuracy:**
The proportion of real outcomes, both true positives and true negatives, within the total number of cases studied is represented by the accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

- **Precision**:
It is the ratio of correctly predicted positive outcomes to all positive outcomes.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

3. Results and Discussion

Dataset is split into two portions named train size and test size. Training and testing in this paper have been applied using Kaggle. During training, the model learns to recognize relationships in between the data, so that it can make accurate predictions on new data. The testing set is used to evaluate the performance of the trained model on new unseen data. Common values for training are 70% and 80%. Therefore, test size is 30% and 20% respectively.

3.1 Experiment 1

Eight different supervised machine learning techniques were applied to the WBCD dataset. Then compared their performance results using five evaluation parameters. Results are summarized in Table 2.

Table 2. Results of experiment 1

Test Size	Model	Accuracy	Precision	F1-Score	Recall	Run Time (ms)
20%	LR	0.973684	1.000000	0.961039	0.925	9.038687
	SVM	0.964912	0.973684	0.948718	0.925	9.028435
	KNN	0.973684	1.000000	0.961039	0.925	2.611399
	DT	0.894737	0.818182	0.857143	0.900	10.607243
	RF	0.956140	0.948718	0.936709	0.925	239.273787
	NB	0.929825	0.900000	0.900000	0.900	3.253937
	XGB	0.938596	0.902439	0.913580	0.925	96.074820
	ADA	0.929825	0.880952	0.902439	0.925	184.181452
30%	LR	0.982456	1.000000	0.973913	0.949153	8.071423
	SVM	0.964912	0.949153	0.949153	0.949153	7.560968
	KNN	0.964912	0.964912	0.948276	0.932203	2.565384
	DT	0.918129	0.846154	0.887097	0.932203	7.824421
	RF	0.947368	0.916667	0.924370	0.932203	225.929499
	NB	0.935673	0.900000	0.907563	0.915254	3.963470
	XGB	0.953216	0.918033	0.933333	0.949153	948.010445
	ADA	0.941520	0.876923	0.919355	0.966102	170.616388

This proposed pipeline model indicates that KNN yielded the highest scores of 97.37% and 100% in terms of accuracy and precision respectively in less computational time than Logistic Regression when the training size is 80%. On the other hand, Logistic Regression outperformed the other machine learning models through a score of 98.24% accuracy and 100% precision when the training size is 70%.

3.2 Experiment 2

The WBCD dataset contains 29 features - excluding "ID' - which is considered a large number of attributes with respect to the size of dataset. Therefore, Feature selection is implemented in this experiment to select the most relevant features to the classification target and exclude the irrelevant features. Feature selection is applied using ANOVA since the attributes are numeric variables. Feature importance scores are shown in Figure 3. Results of selecting different number of attributes are summarized in Table 3.

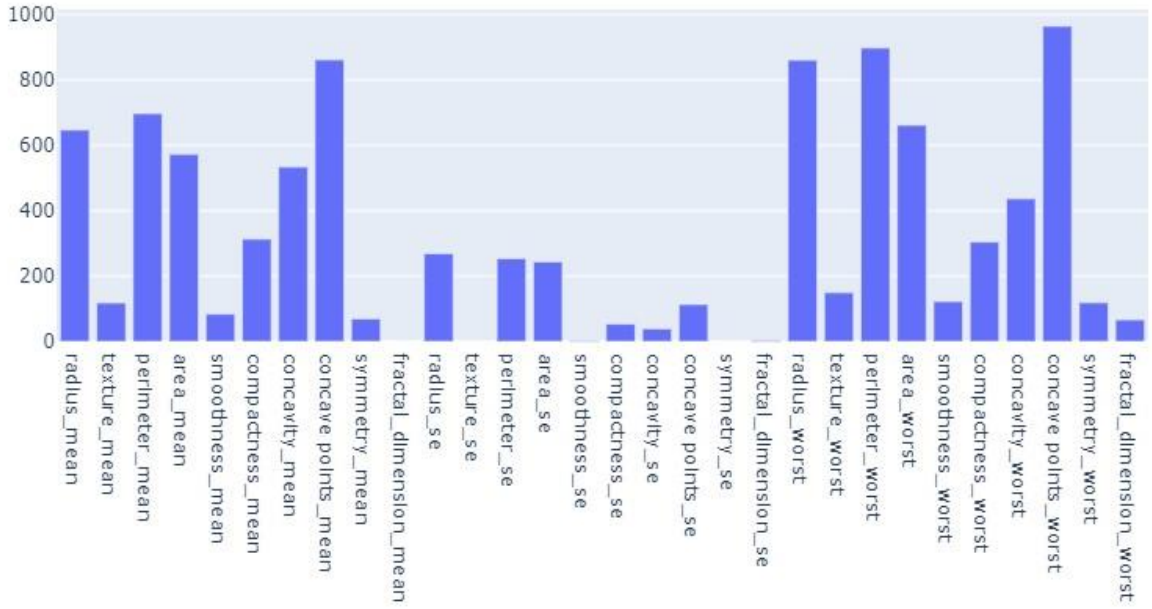

Table 3. Results of experiment 2

No. of features	Model	Accuracy	Precision	F1-Score	Recall	Run Time (ms)
22	LR	0.982456	1.000000	0.973913	0.949153	6.593943
	SVM	0.959064	0.948276	0.940171	0.932203	8.327961
	KNN	0.964912	0.949153	0.949153	0.949153	3.349543
	DT	0.883041	0.809524	0.836066	0.864407	7.284641
	RF	0.947368	0.903226	0.925620	0.949153	283.756256
	NB	0.923977	0.870968	0.892562	0.915254	2.994061
	XGB	0.953216	0.918033	0.933333	0.949153	69.896221
	ADA	0.941520	0.888889	0.918033	0.949153	171.515226
19	LR	0.982456	1.000000	0.973913	0.949153	6.488562
	SVM	0.970760	0.965517	0.957265	0.949153	5.947828
	KNN	0.970760	0.965517	0.957265	0.949153	2.944231
	DT	0.894737	0.805970	0.857143	0.915254	6.383419
	RF	0.941520	0.888889	0.918033	0.949153	274.549246
	NB	0.929825	0.885246	0.900000	0.915254	3.957748
	XGB	0.947368	0.916667	0.924370	0.932203	65.920353
	ADA	0.935673	0.875000	0.910569	0.949153	161.902666
15	LR	0.935673	0.900000	0.907563	0.915254	5.768061
	SVM	0.935673	0.900000	0.907563	0.915254	5.643368
	KNN	0.929825	0.885246	0.900000	0.915254	3.302813
	DT	0.912281	0.833333	0.880000	0.932203	5.343676
	RF	0.935673	0.875000	0.910569	0.949153	266.089439
	NB	0.918129	0.868852	0.883333	0.898305	3.699064
	XGB	0.941520	0.888889	0.918033	0.949153	64.128876
	ADA	0.929825	0.861538	0.903226	0.949153	155.796766
12	LR	0.941520	0.901639	0.916667	0.932203	4.546642
	SVM	0.953216	0.947368	0.931034	0.915254	5.355597
	KNN	0.953216	0.932203	0.932203	0.932203	2.788782
	DT	0.929825	0.873016	0.901639	0.932203	4.557610
	RF	0.941520	0.901639	0.916667	0.932203	289.741278
	NB	0.923977	0.870968	0.892562	0.915254	2.791643
	XGB	0.941520	0.901639	0.916667	0.932203	63.071251
	ADA	0.935673	0.887097	0.909091	0.932203	152.271032

This proposed method indicates that LR kept the highest performance scores and 98.24% in terms of accuracy while using a fewer number of 22 and 19 predictive features out of total number of 30 attributes. Extreme Gradient Boosting scored an accuracy of 94.15% when implementing half number of features only. KNN reached an accuracy score of 95.32% when 12 features were included in prediction model which is considered about third of total number of predictive attributes in a less running time than Support Vector Machine.

3.3 Experiment 3

Large-scale medical datasets can be predicted and classified with greater accuracy using neural networks. In this experiment, Different parameters shown in Table 4 of the Neural network is hyper-tuned to achieve the best results. Accuracy and Test time shown in table 5. While accuracy and loss versus epochs is shown in Figure 4 at different number of epochs.

Table 4. Neural Network Parameters

Batch Size	16
Learning Rate	0.0001
Activation Function	Sigmoid
Optimizer	Adams Solver

Table 5. Results of experiment 3

No. of epochs	Accuracy	Time in S
80	97.08	7.25
100	97.66	8.24
120	98.25	10.02
150	97.66	21.13

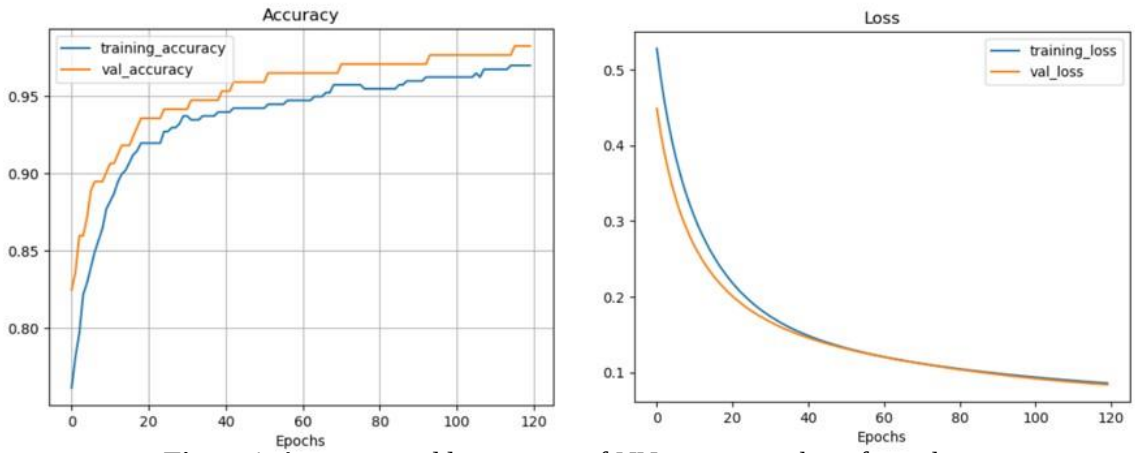

Figure4. Accuracy and loss curves of NN versus number of epochs

The neural network reached an accuracy of 98.25% at 120 epochs. But when increasing number of epochs, the accuracy remained at 97.66% as well as in 100 epochs but there was a significant increase in running time and computational cost.

4. CONCLUSION

Nowadays, the accurate prediction of breast cancer disease considered to be one of the challenging medical research topics. This research has included in the deployment of a ML-based pipeline to successfully classify breast cancer using a data set of 569 instants and 30 features. Consequently, our goal was met. by utilizing and analyzing various ML algorithms such as Random Forest, Decision Tree, AdaBoost, K-Nearest Neighbors, Xgboost, and Gaussian Naive Bayes as well as artificial neural network, then compared the performance of these algorithms.

The proposed model demonstrated an accuracy of 98.25% when applying feature selection methods through Logistic Resgression as well as neural network. Validation and testing were performed. In future work, other advanced Machine Learning and Deep Learning techniques will be verified on a different type of medical datasets images so that the efficiency and effectiveness of proposed model prediction can be enhanced at earlier stages.

REFERENCES

[1] M. Abdar *et al.*, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognit Lett*, vol. 132, pp. 123–131, 2020, doi: https://doi.org/10.1016/j.patrec.2018.11.004.

[2] Q. H. Nguyen *et al.*, "Breast Cancer Prediction using Feature Selection and Ensemble Voting," in *2019 International Conference on System Science and Engineering (ICSSE)*, IEEE, Jul. 2019, pp. 250–254. doi: 10.1109/ICSSE.2019.8823106.

[3] V. Chaurasia and S. Pal, "Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer," *SN Comput Sci*, vol. 2, no. 2, p. 67, Apr. 2021, doi: 10.1007/s42979-021-00465-3.

[4] M. R. Basunia, I. A. Pervin, M. Al Mahmud, S. Saha, and M. Arifuzzaman, "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.

[5] M. A. Naji, S. El Filali, K. Aarika, E. L. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Comput Sci*, vol. 191, pp. 487–492, 2021, doi: https://doi.org/10.1016/j.procs.2021.07.062.

[6] K. Jabeen *et al.*, "BC2NetRF: Breast Cancer Classification from Mammogram Images Using Enhanced Deep Learning Features and Equilibrium-Jaya Controlled Regula Falsi-Based Features Selection," *Diagnostics*, vol. 13, no. 7, p. 1238, Mar. 2023, doi: 10.3390/diagnostics13071238.

[7] L. K. Singh, M. Khanna, and R. Singh, "Artificial intelligence based medical decision support system for early and accurate breast cancer prediction," *Advances in Engineering Software*, vol. 175, p. 103338, 2023, doi: https://doi.org/10.1016/j.advengsoft.2022.103338.

[8] A. Al Bataineh, "A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection," *Int J Mach Learn Comput*, vol. 9, pp. 248–254, Jun. 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[9] M. Mangukiya, A. Vaghani, and M. Savani, *Breast Cancer Detection with Machine Learning*, vol. 10. 2022. doi: 10.22214/ijraset.2022.40204.

[10] S. Y. Yashfi *et al.*, "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–5. doi: 10.1109/ICCCNT49239.2020.9225548.

[11] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J Supercomput*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.

[12] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Appl Soft Comput*, vol. 88, p. 105946, Mar. 2020, doi: 10.1016/j.asoc.2019.105946.

[13] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J Clin Epidemiol*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[14] A. B. Møller, B. V. Iversen, A. Beucher, and M. H. Greve, "Prediction of soil drainage classes in Denmark by means of decision tree classification," *Geoderma*, vol. 352, 2019, doi: 10.1016/j.geoderma.2017.10.015.

[15] J. Speiser, M. Miller, J. Tooze, and E. Ip, "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling," *Expert Syst Appl*, vol. 134, May 2019, doi: 10.1016/j.eswa.2019.05.028.

[16] M. Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," Jul. 2016, pp. 35–39. doi: 10.1109/DeSE.2016.8.

[17] M. Desai and M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)," *Clinical eHealth*, vol. 4, pp. 1–11, 2021, doi: https://doi.org/10.1016/j.ceh.2020.11.002.

[18] A. Ogunleye and Q. G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 17, no. 6, 2020, doi: 10.1109/TCBB.2019.2911071.

[19] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.