Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 11 Issue 2, February - 2024

Breast Cancer Detection Using Logistic Regression And Random Forest Machine Learning Techniques

Esther. M. Umoren[1]
Department of Computer Engineering,
University of Uyo, Akwa Ibom, Nigeria
SS

Philip Michael Asuquo[2]
Department of Computer Engineering,
University of Uyo, Akwa Ibom, Nigeria
asuquophhilip@gmail.com

Stephen, Bliss Utibe-Abasi[3]
Department of Computer Engineering,
University of Uyo, Akwa Ibom, Nigeria

***Abstract*— The aim of this study is to determine from a patient's medical history whether breast cancer is benign or malignant using logistic regression and random forest machine learning techniques. The dataset was obtained from Breast Cancer Wisconsin (Diagnostic) dataset repository. In the dataset, there are 35 attributes and 569 patient records. The 5-fold cross-validation was adopted whereby 20% of the dataset was used as the test set while the other 80% of the fold was used for the training of the models. After training, it was discovered from the model performance results that the logistic regression model performed better than the random forest model. Specifically, the logistic regression model has the best performing with a 98.7% training accuracy, 96.7% validation accuracy, and a 97.2% test accuracy. In addition, the logistic regression has the highest test F1-score of 97.163% and the highest recall score of 96.92% which is among the most important metric for the Healthcare industry. The ideas presented in this work will help doctors to quickly detect breast cancer and this can reduce maternal mortality due to breast cancer.**

***Keywords*— Breast Cancer, Machine Learning, Classification, Logistic Regression , Confusion Matrix, K-Nearest Neighbor, Random Forest**

1. INTRODUCTION

According to research findings, when cells begin to multiply uncontrollably, cancer develops [1,2,3,4]. Cancer in the breast is a condition that is identified by the irregular changes of breast cells. It is the most prevalent malignant disorder that affects women and the main reason why they die [9,10,11]. Cancerous cells associated with breast cancer carriages a great risk to health and contributes significantly to female fatality [12,13,14]. The number of breast cancer cases has intensified over the years, due to the disease's increasing global growth amongst women, it has risen to the status of the second-leading sickness [32,33,37]. The number of death due to breast cancer can be actually reduced with an early diagnosis. Since it is one of the most curable tumors if diagnosed earlier, medical management improves the chances of survivability.

The incidence of cancerous cells forming in the breast, is more common in females than men. Age, family history, and medical history are a few of the most significant risk variables that you have no control over [21,22,23]. The incidence of cancerous cells forming in the breast, is more common in females than men [15,24,]. Family history and medical history are a few of the most significant risk variables that you have no control over [7,8,17]. Breast cancer diagnosis is a lengthy and costly process. The consistency and expertise of the medical experts are crucial to the diagnosis procedure [,25,26,27]. Malignant and benign tumors are the two primary types of growths (tumorous). Diagnosis by medical professionals is frequently prone to omissions and errors, therefore this misinterpretation might send the potential breast cancer patient into irreversible damage [5,6,31].

Meanwhile, nowadays, information gathering can now be done with incredible speed, scale, and complexity while needing less work and having more autonomy [28,29,30]. In order to execute a variety of activities, data are rapidly being merged with developing technologies on various sizes, from cellphones to computing, as the information and experience in these endeavors increase [34,35,36]. Data mining and machine learning, two overlapping subfields of artificial intelligence, examine the connections between determinants and consequences, forecast future issues, and offer remedies to some life problems [39,40,16]. Accordingly, the objectives of this study is the application of two different machine learning models in the detection of breast cancer using a dataset

along with 5-fold cross validation method to train and evaluate the performance of the models [18,19,20].

2. RELATED WORKS

Many researchers have carried out research on the prediction of breast cancer using the various Machine Learning algorithms. These researchers used relevant datasets, carried out exploration of the dataset, as well as extraction and selection of various features that could aid in the completion of the research. Shwetha, et al., in [41] made use of the Convolution Neural network (CNN) to detect breast cancer, CNN is a deep learning technique that is used for analyzing visual images. It is also a machine learning tool that has much importance in the visual detection of health-related problems. The authors used mammogram image from the dataset. The model was trained using the Mammographic Image Analysis Society (MIAS) database and tested, the results were obtained using CNN classifier which showed the accuracy and validity of the model. Also, Khali et al., in [42] made use of WEKA (Waikato Environment for Knowledge Analysis) for data mining and execution of tasks such as image and text classification. The authors discovered that the J48 had a better percentage accuracy compared to other algorithms, while Random Forest Tree had the lowest. Also, Salva and Kadam in [43] made use of CNN. They made use of a dataset consisting of images of preciously encountered cancer cells in the region of the breast of the patient diagnosed with breast cancer. To train the model, they divided the dataset into 80% training data, 10% validation data, and 10% testing data. The performance metric of the analysis showed that the model produced an accuracy of 87.84% and a precision of 75.3%. This research aim at analysing two machine learning algorithms to figure out their effectiveness, strength and weaknesses in breast cancer detection and to guide future researchers in selecting the most suitable approach.

3. METHODOLOGY

The data used for this research is obtained from Wisconsin breast cancer dataset (Kaggle repository). The dataset contains records of 569 patients and 35 columns. After the data was obtained, a pre-processing was done to check for of any missing values in the dataset. The features were separated from the target variables and kept on the same scale by applying standardization. This keeps the values of each column between the range of -3 and 3. It also makes the model training process faster. The dataset is split into five folds, and the model is evaluated on each fold of the data. The model is created by fitting the two different machine learning algorithms to the cleaned data. The diagram in Figure 1, shows the process with which the dataset is analyzed.

The machine learning algorithms that are used in this work are logistic regression and random forest. The performance metric of the models includes accuracy, precision, recall and f1-score. The hyper parameters of the algorithms used in this work are fine-tuned using a technique called GridSearchCV which is meant to give the highest F1-Score. The dataset used for this work are features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

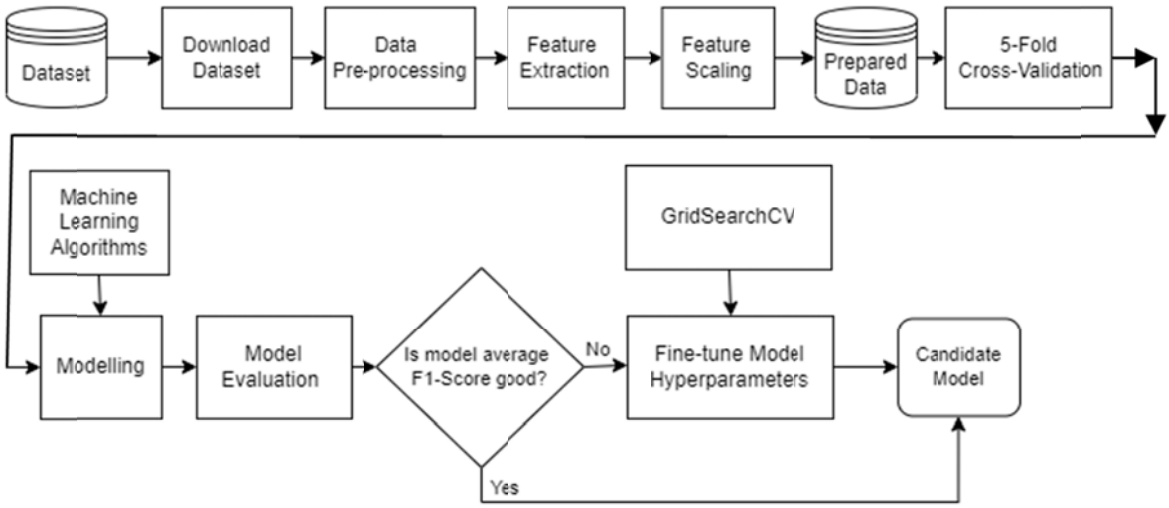


Figure 1. Flowchart of the research model

3.1 Training with the Logistic Regression Algorithm

LogisticRegression class was imported from the *linear_model* module of the Sci-kit Learn library. GridSearchCV was used for hyper-parameter tuning. It is employed to find the best hyperparameters combination that gives the highest f1-score on the validation set. The regularization parameter *C* is 3.728. The regularization parameter is used to minimize overfitting of the model. The maximum number of training iterations, *max_iter* was set to 2000 as shown in Figure 2.

```
# Use the best parameters to train the logistic regression algorithm
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression

lr_model = LogisticRegression(C=3.727593720314938, penalty='l2', max_iter=2000)
lr_train_val_result = cross_validation(lr_model, scaled_X_train, y_train, 5)
print(lr_train_val_result)
```

Figure 3.14: Training with the Logistic Regression Algorithm

3.2 Training with the Random Forest Algorithm

RandomForestClassifier class was Imported from the *ensemble* module of the Sci-kit Learn library. The *n_estimators* parameter represents the number of decision trees we use to build our random forest classifier, as can be seen in Figure 3.

The *n_estimators* parameter represents the number of decision trees we use to build our random forest classifier. The *n_estimators* parameter is 130. It means we are using 130 decision trees in our random forest classifier. The *min_samples_split* is 10, and the *criterion* parameter is '*entropy*.'

```
# Use the best parameters to train the Random Forest algorithm
rf_model = RandomForestClassifier(n_estimators=130,
                                  min_samples_split=10,
                                  criterion="entropy",
                                  random_state=0)

rf_train_val_result = cross_validation(rf_model, scaled_X_train, y_train, 5)
print(rf_train_val_result)
```

Figure 3. Training with the Random Forest Algorithm

4. RESULTS AND DISCUSSION

The 5-fold approach is used in the splitting of the dataset for training and testing. In this case, 80 % of the dataset is used in the model training while 20% of the data is kept aside for use as a test set. The F1 score metric results of each of the two models are shown in Table 1 and Figure 4 for the training dataset and in Table 2 and Figure 5 for the test set. Again, the accuracy, precision, recall, and f1-score metrics was used to evaluate and compare the performance of the two models. The results on those performance metrics for the two models are shown in Table 3 and Figure 6.

Based on the results in Table 1 to Table 3, along with the same results presented in Figure 3 to Figure 6, the best performing model is the Logistic Regression model. It has the highest F1-score of 97.158% and the highest accuracy score of 97.368%. It also has the highest recall score of 96.925%. On the other hand, the Random Forest model has F1-score of 96.149 %, accuracy score of 96.49% and recall score of 95.238%.

Table 1: Training F1 scores across the 5 folds

	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
Logistic Regression	98.8848	98.1413	99.6310	98.1273	98.8848	98.7338
Random Forest	98.5075	98.8848	99.2647	97.7778	98.5185	98.5906

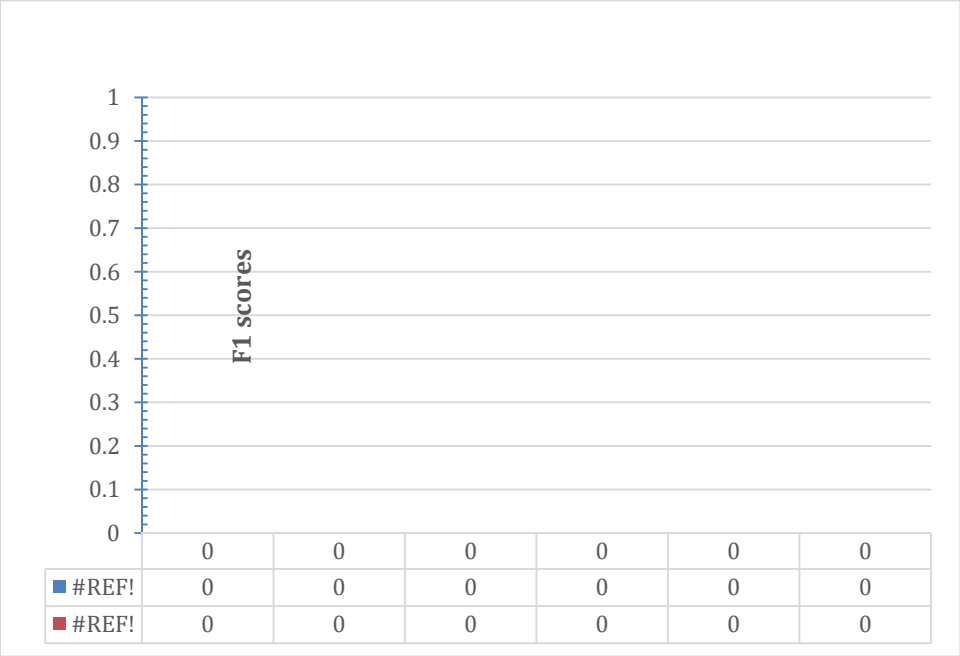


Figure 4 The F1 scores across the 5 folds for the training dataset

Table 2: Validation F1 scores across the 5 folds

	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
Logistic Regression	96.9697	98.5507	93.7500	100.0000	94.1176	96.6776
Random Forest	95.3846	100.0000	89.2308	95.5224	88.5714	93.7418

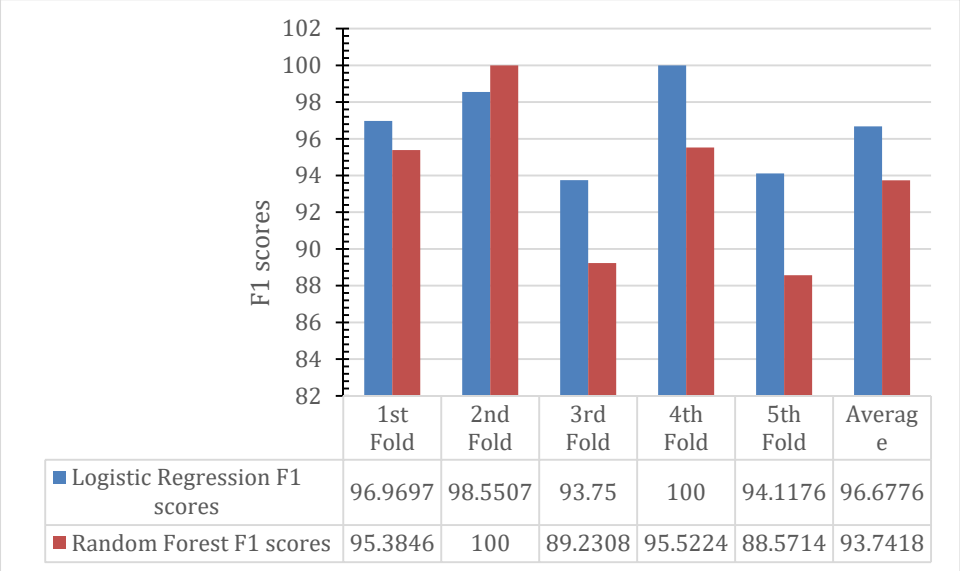
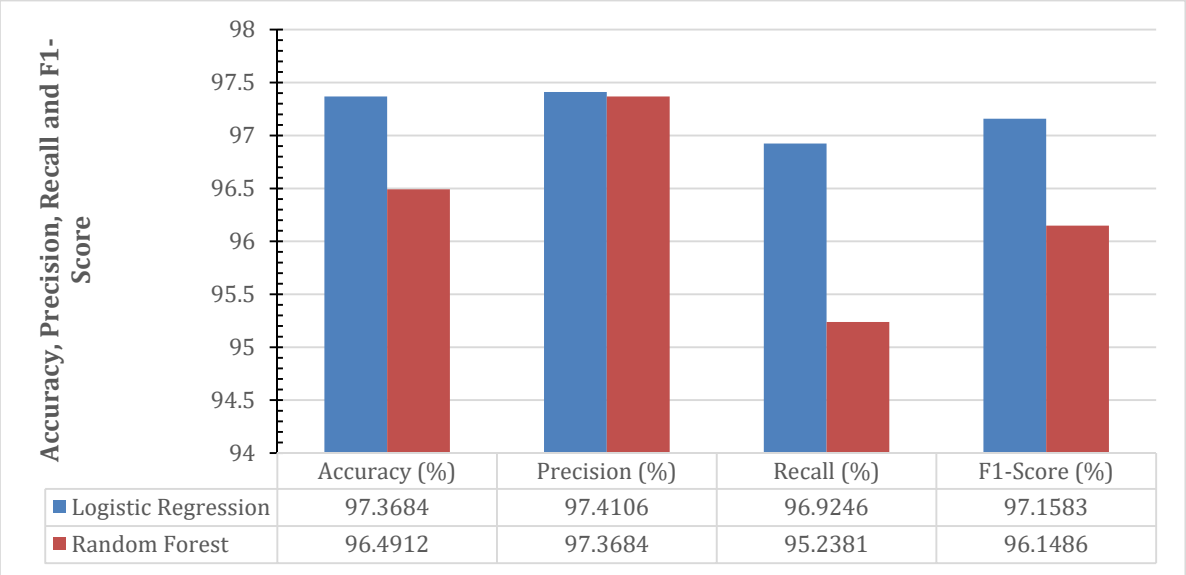


Figure 5 The F1 scores across the 5 folds for the validation dataset

Table 3: The Results on Accuracy, Precision, Recall and F1-Score

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	97.3684	97.4106	96.9246	97.1583
Random Forest	96.4912	97.3684	95.2381	96.1486



	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	97.3684	97.4106	96.9246	97.1583
Random Forest	96.4912	97.3684	95.2381	96.1486

Figure 5 The Results on Accuracy, Precision, Recall and F1-Score

6. CONCLUSION

This work focused on using logistic regression and random forest machine learning techniques to diagnose breast cancer from patients' medical records. By evaluating and comparing the performance of the two algorithms on the Wisconsin breast cancer dataset, new insights into their effectiveness in diagnosing breast cancer are obtained. The 5-fold cross-validation was used to evaluate the performance of the model. The validation accuracy of all the models was above 90%. Logistic regression model has better performance. It has the highest f1-score (97.16%). The validation accuracy of all the models was above 90%. The best performing model, the logistic regression model, had an accuracy of 97%. Given that the logistic regression model achieved the highest recall score it will be the preferred model for the health sector. Healthcare professionals can benefit from this knowledge when developing and utilizing breast cancer detection systems. Researchers and healthcare practitioners can rely on this information to make informed decisions when choosing an algorithm for breast cancer detection.

REFERENCE

1. Wang Z, Li, M., Wang, H., Jiang H., Yao Y. (2019). "Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features," *IEEE Access*, vol. 7, no. 7, pp. 105146–105158.
2. Gladju J., Kamalam B. S. and Kannagaraj A. (2022) Applications of data mining and machine learning framework in aquaculture and fisheries: review. Smart Agricultural Technology 2.: 1-15
3. Sarker I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* 2: 1- 39
4. Jiang Y, Chen L., Zhang H. and Xiao X. (2019). "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module," *PLOS ONE*, vol. 14, no. 3, pp. 1–21.
5. Thomas D. and Ravi K. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal. 6(2): 94 – 98
6. Fondón I., A., Sarmiento A., García A. I. Silvestre M. and Eloy C. (2018). "Automatic classification of tissue malignancy for breast carcinoma diagnosis," *Computers in Biology and Medicine*, vol. 96, pp. 41–51
7. Gladju J., Kamalam B. S. and Kannagaraj A. (2022) Applications of data mining and machine learning framework in aquaculture and fisheries: review. Smart Agricultural Technology 2.: 1-15
8. Kutrani H. and Eltalhi S. (2019). Breast cancer diagnosis and prediction using machine learning and data mining techniques: a review. *Journal of Dental and Medical Sciences (IOSR-JDMS)*, 18(4): 1- 7
9. Saad A. A., Kamruzzaman M. M., Sarker M. N., Alruwaili M., Alhwaiti Y., Alshammari N. and Siddiqi M. H. (2021). Boosting Breast Cancer Detection Using Convolutional Neural Network. *Journal of Healthcare Engineering* (2021): 1-7
10. Olofintuyi S. S. (2023). Breast Cancer Detection with Machine Learning Approach. *FUDMA Journal of Sciences (FJS)*, 7(2): 1-17
11. PATTANAIK R., GELMECHA D. J., SIDDIQUE M., SINGH R. S., SATAPATHY S., AND MISHRA R. S. (2023). BREAST CANCER DETECTION AND CLASSIFICATION USING METAHEURISTIC OPTIMIZED ENSEMBLE EXTREME LEARNING MACHINE. *INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY:* 15: 4551- 4563
12. KAMIL M.Y. AND SALIH A. M. (2019). MAMMOGRAPHY IMAGES SEGMENTATION VIA FUZZY C-MEAN AND K-MEAN. *INTERNATIONAL JOURNAL OF INTELLIGENT ENGINEERING AND SYSTEMS*, 12(1): 1-8
13. Mishra A.K., Roy P., Bandyopadhyay S. and Das S. K. (2022). Achieving highly efficient breast ultrasound tumor classification with deep convolutional neural networks. *International*

Journal of Information Technology, 14:3311–3320.

14. Madallah A. and Walaa (2022). Automated breast cancer detection models based on transfer learning. *Sensors,* 22: 1-16
15. Adam R., Dell'Aquila K., Hodges L., Maldijan T. and Duong T. (2023). Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review. *Breast Cancer Research,* 25: 1-12
16. Nassif A., Talib M. A., Nasir Q., Afadar Y. and Elgendy O. (2022). Breast cancer detection using artificial intelligence techniques: a systematic literature review. *Artificial Intelligence in Medicine:* 1-26
17. Maha M. A., Amal Adnan A., Nada A. A., Asim A., Romany F. M., Deepak G. and Ashish K. (2022). Deep transfer learning-based breast cancer detection and classification model using photoacoustic multimodal images. *BioMed Research International*: 1-13
18. Jaehoon L., Hee S. L., Soo B. P., Chanyang K., Kahee K., Dawoon E., and Si Y. (2021). Identification of circulating serum miRNAs as novel biomarkers in pancreatic cancer using a penalized algorithm. *International Journal of Molecular Sciences,* 22: 1-13
19. Pattnaik R. K., Siddique M., Mishra S., Gelmecha D. J., Singh R. S. and Stapathy S. (2023). Breast cancer detection and classification using metaheuristic optimized ensemble extreme learning machine. *International Journal of Information Technology*: 1-14
20. Shin-Yi C., Ta-Chung C. and Yeu S. (2020). The stemness-high human colorectal cancer cells Promote Angiogenesis by Producing Higher Amounts of Angiogenic Cytokines via Activation of the Egfr/Akt/Nf-κB Pathway. *International Journal of Molecular Sciences,* 22: 1-20
21. Rong W. and Yong W. (2021). Fourier transform infrared spectroscopy in oral cancer diagnosis. *International Journal of Molecular Sciences,* 22: 1-21
22. María G., Miguel G., Alicia R., Beatriz A., Diego M., Amalia M., Ana F. and Apolonia N. (2021). Genetic Variants of ANGPT1, CD39, FGF2 and MMP9 linked to clinical outcome of bevacizumab plus chemotherapy for metastatic colorectal cancer. *International Journal of Molecular Sciences,* 22: 1-16
23. Farhat A., Muhammad S., Muhammad A. K., Usman T., Hwan-Seung Y. and Jaehyuk C. (2022). Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. *Sensor*, 22: 1-22
24. Yanyu X., Nantao L., Congnyu C., Weijing W., Priyash B., Weinan L., Leyang L., Xiaojing W., Shaoxiong W., Huan H. and Brian T. C. (2022). Microscopies enabled by photonic metamaterials. *Sensor*, 22: 1-28
25. Ranpreet K., Hamid Gholam H., Roopak S. and Maria L. (2022). Melanoma classification using a novel deep convolutional neural Network with dermoscopic images. *Sensor*, 22: 1-15
26. Himanish S. D., Akalpita D., Anupal N., Saurav M., Kangkana B. and Zhongming Z. (2022). Breast cancer detection: Shallow convolutional neural network against deep convolutional neural network based approach. *Frontier in Genetics*: 1-14
27. Janice T., Jessica P., Bryan V. and Marie-Paule G. (2021). Hyperglycemic condition causes pro-inflammatory and permeability alterations associated with monocyte recruitment and deregulated nfκb/pparγ pathways on cerebral endothelial cells: Evidence for Polyphenols Uptake and Protective Effect. *International Journal of Molecular Science*: 1-19
28. Kiran J., Muhammad A. K., Majed A., Usman T., Yu-Dong Z., Ameer H., Arturas M. and Robertas D. (2022). Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*: 1-23
29. Deepraj C., Anik D., Ajoy D., Shreya S., Ashutosh D. D., Raghava R. M. and Lakhindar M. (2021). ABCanDroid: A cloud integrated android app for noninvasive early breast cancer detection using transfer learning. *Sensors*: 1-20
30. Xiao J., Xiaolin S., and Xingang Z. (2022). Breast cancer identification using machine learning. *Mathematical Problems in Engineering*: 1-8
31. Zhe M. and Yang L. (2023). Nomogram based on super-resolution ultrasound images outperforms in predicting benign and malignant breast lesions. *Breast Cancer: Targets and Theraphy* 15: 867-878
32. Marinovich M. L., Elizabeth W., William L.,4 Alison P., Stacy M C., Helen L.. Andrew W., Jiye G. K., Gavin F. P.,1,6 Christoph I. L., Sophia Z., Meagan B. Nehmat H. (2022). Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. *BMJ Open*: 1-8
33. Hameedur R., Tanvir F. N. B., Rozilawati A., Ahmad A. and Abdul R. J. (2023). Efficient breast cancer diagnosis from complex mammographic images using deep convolutional neural network. *Computational Intelligence and Neuroscience*: 1-11
34. Mariam Raafat1 , Sahar Mansour1* , Rasha Kamal1 , Hedaya W. Ali1 , Passant E. Shibel2 , Ahmed Marey3 , Sherif N. Taha4 and Basma Alkalaawy (2022). Does artifcial intelligence aid in the detection of diferent types of breast cancer. *Egyptian Journal of Radiology and Nuclear Medicine,* 53(1): 1-9
35. Basem S. A., Mohammed R. J. A., Ihab S. Z. and Samy S. A. (2023). Convolution neural network for breast cancer detection and classification using deep learning. *Asian Pacific Journal of Cancer Prevention*, 24(2): 1-14
36. Memon Q. and Asadi B. (2023). Efficient breast cancer detection via cascade deep learning

network. *International Journal of Intelligent Networks,* 4: 46- 52

37. Putri S. R. S. and Nuzula R. (2022). A scoping review: accuracy in early detection of breast cancer using a clinical breast examination method. *Journal of Public Health Science* (JPHS), 1(1): 1-8
38. Li S., Laurie R. M., Joseph H. R., Eugene F., Russell M. and Weiva S. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports* :1-12
39. Francesco P., Marco I., Alessia O., Salvatore G. and Salvatore V. (2023). A yolo-based model for breast cancer detection in mammograms. *Cognitive Computation*: 1-14
40. Basem S. A., Mohammed R. J. A., Ihab S. Z. and Samy S. A. (2022). Breast cancer detection and classification using deep learning xception algorithm. International Journal of Advanced Computer Science and Applications (IJACSA) 8: 1-6
41. Shwetha K., Spoorthi M., Sindhu S. S., Chaithra D., (2018). Breast cancer detection using deep learning technique. *NCESC,* 6 (13): 1-8
42. Khalil A., Ahmed I., Khan Z. H., Siddiqui S. I. and Ahmad I. (2020). Machine Learning Algorithms for Early Stage Breast Cancer Diagnosis, *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9* (09): 369-372
43. Salvi S. and Kadam A. (2020). Breast Cancer Detection Using Deep Learning and IoT Technologies. Journal of Physics Conference Series 1831(1): 1-8