iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38582

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Breast Cancer Detection using Artificial Neural Networks

Md Haris Uddin Sharif[1]
[1]*Department of Information Technology, University of the Cumberlands*

Abstract: Early detection of disease has emerged as a critical issue in recent years due to the fast population increase seen in medical research. The chance of dying from breast cancer increases dramatically as the world's population continues to increase at an alarming rate. Compared to other cancers discovered thus far, breast cancer is the second most severe. In addition to assisting medical staff in disease diagnosis, an automated disease detection system also provides reliable, effective, and fast intervention, which reduces the likelihood of mortality. In this research study, the Artificial Neural Network is employed for breast cancer classification. The model is validated on well-known dataset comprised from UCI machine learning repository. The results reveal that the ANNs obtained the highest accuracy i.e. 98.24%.

Keywords: Machine Learning, Neural Network, Algorithm, Artificial Intelligence.

I. INTRODUCTION

The correct identification of some essential information is a significant problem in bioinformatics and medical research, among other fields [1]. In the field of medicine, the diagnosis of a disease is a physically demanding and challenging task. Thousands of diagnostic centres, hospitals, and research institutes, in addition to countless websites, provide a wealth of medical diagnosis information to the public at large. It is scarcely required to categorize them to make the system automated and fast in identifying medical conditions. The expertise and ability of the medical planning officer in the medical field are often used to diagnose a requirement in most cases. Consequently, there are situations in which mistakes and undesirable biases occur. It also takes a long time to get an accurate diagnosis of the illness.

According to the American Cancer Society [2], women are more likely than men to be affected with breast cancer than all other cancers discovered. Approximately one-third of the female population is infected with invasive breast cancer, according to estimations. Breast cancer is the most common kind of cancer in women all over the world. Breast cancer develops as a result of the abnormal development of specific cells inside the breast. Several methods have been developed to ensure that breast cancer is diagnosed correctly. Breast screening, often known as mammography [3], is used to detect and diagnose breast cancer. By using X-rays, it is possible to determine the nipple status of a woman. Breast cancer is difficult to detect in its early stages in most cases, owing to the tiny size of the cancer cell when seen from the outside. It is possible to detect cancer in its early stages using mammography, and the procedure takes just a few minutes. When it comes to detecting breast cancer, ultrasound [4] is a well-known method in which a sound wave is delivered into the body to examine the situation on the inside. A transducer that provides sound waves is located on the skin, and the bounce of the sound waves records the echoes of the tissues of the body as they travel through the body. It is required to convert the echoes into a greyscale, a binary value represented in a computer.

Positron emission tomography (PET) [5], which uses F-fluorodeoxyglucose to image the human body, allows physicians to determine the location of a tumour in the body. It is based on the detection of radiolabel cancer cells in the body. -tracers that are specific. Breast distortions may be detected using dynamic magnetic resonance imaging (MRI), which has been developed [6]. The modality predicts the pace of contrast enhancement in cancer by increasing the rate of angiogenesis in the cancerous tissue. The presence of metastases on magnetic reasoning imaging in breast cancer patients is associated with increased contrast enhancement. As a consequence of advancements in imaging technology, the method known as Elastography [7] has just been created. Breast cancer tissue that is larger than the normal parenchyma may be removed using this method. This method uses a colour map of probe compression to distinguish between benign and malignant tumours.

Medical prognosis has significantly benefited from the application of machine learning [8–11], deep learning [12, 13], and bio-inspired computing [14]. There have been many methods shown, but none of them have provided an accurate and reliable result. Doctors must interpret a large amount of imaging data during mammography, which decreases accuracy. This technique is highly time-consuming, and in some cases, it incorrectly diagnoses the illness. This paper proposed a machine learning-based technique (Artificial Neural Networks) to detect the disease from the input features.

The remaining part of the paper is prepared as follows. The following section outlines the current review of state of the art in this field, followed by which the methods and materials used for the study are illustrated. Section III presented the details of the dataset, and section IV presented the experimental protocol, while section V concluded the paper.

II. RELATED WORK

Many innovative systems for detecting breast cancer have been created as medical science has progressed. The following is a survey of the research in this area: Sakri et al. [15] utilized a feature selection method called particle swarm optimization (PSO) in conjunction with machine learning algorithms K-NNs, Naive Bayes (NB), and the reduced error pruning (REP) tree to improve the accuracy value. According to their study, Saudi Arabian women's breast cancer is one of its main issues. According to their findings, this disease primarily affects women over the age of 46.

Keeping this in mind, the authors of [15] used the WBCD dataset to test four phase-based data processing methods. They published a study that compared classification without a feature selection technique to category with a feature selection method. For NB, RepTree, and K-NNs, they achieved 70 percent, 76.3 percent, and 66.3 percent accuracy, respectively. They utilized the Weka tool to do their data analysis. They discovered four characteristics that are optimal for this classification job after using PSO. They achieved accuracy values of 81.3%, 80%, and 75%for NB, RepTree, and K-NNs using PSO, respectively.

Kapil and Rana [16] presented a weight-enhanced decision tree as a modified decision tree method and applied it on WBCD and another breast cancer dataset obtained from the UCI library. They discovered that they have rated each feature and retained the essential characteristics for this classification assignment using the Chi-square test. Their suggested method achieved about 99% accuracy on the WBCD dataset, whereas it achieved roughly 85–90% accuracy on the breast cancer dataset.

On the benchmark, Wisconsin Breast Cancer Diagnosis (WBCD) dataset, Yue et al. [17] primarily showed thorough reviews on SVM, K-NNs, ANNs, and Decision Tree techniques in the application of predicting breast cancer. The authors claim that combining deep belief networks (DBNs) with ANN architecture (DBNs-ANNs) yielded a more accurate outcome. This architecture achieved 99.68% accuracy, while the SVM approach produced 99.10 % classification accuracy using a two-step clustering algorithm and the SVM methodology. They also looked at the ensemble method, which used the voting technique to construct SVM, Naive Bayes, and J48. The accuracy of the ensemble technique was 97.13%.

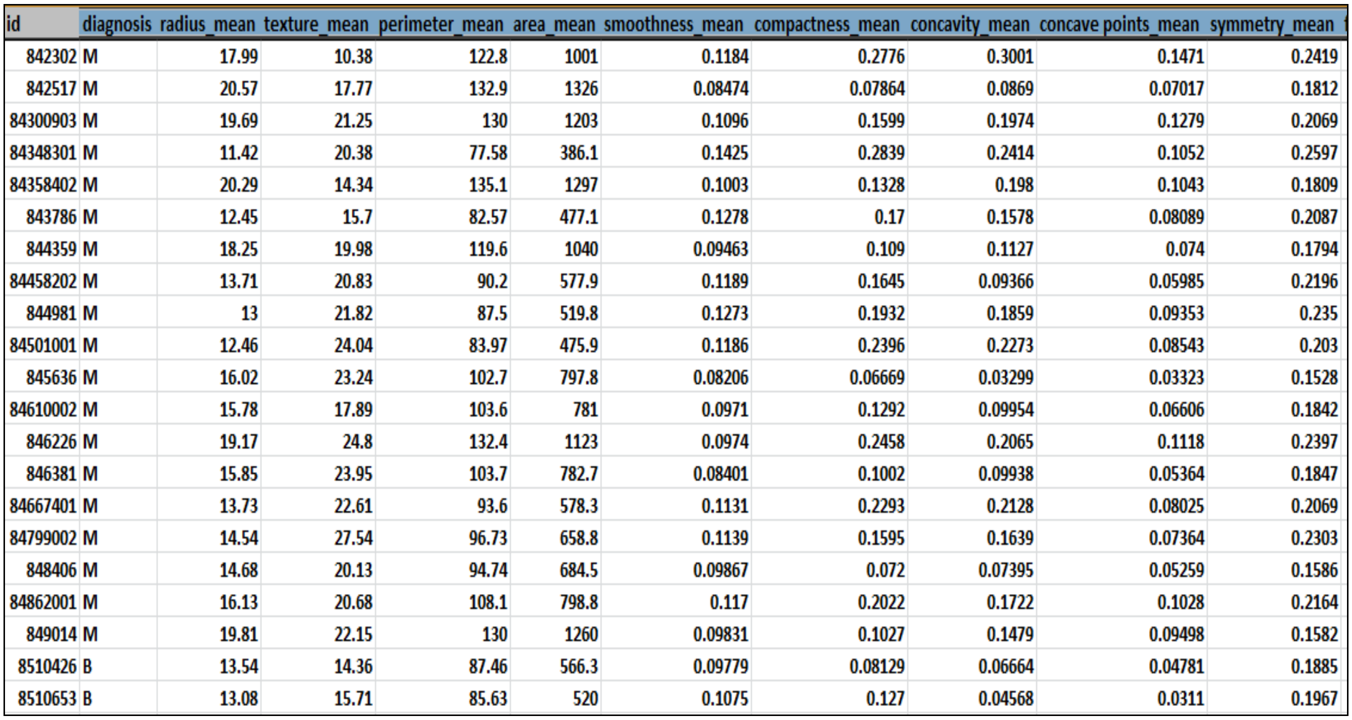
Azar et al. [18] used decision tree variations to develop a technique for predicting breast cancer. A single decision tree, a boosted decision tree, and a decision tree forest are all modalities utilized in this approach (DTF).To arrive at a judgment, a data set must first be trained, followed by testing. In the training phase, there were 97.07 percent and 98.83 percent accuracy results produced by SDT and BDT, respectively, indicating that BDT performed better. Decision tree forest was 97.51% accurate, whereas SDT was 95.75% accurate throughout testing. Ten-fold cross-validation was used to train the dataset.

Breast cancer detection was shown in [19] by the authors. Local linear wavelet neural network (LLWNN) and recursive least square (RLS) were used to identify the disease in this study, which improves the system's performance. The LLWNN-RLS has the highest average Correct Classification Rate (CCR) of 0.897 and 0.972 for two and three predictors, respectively, with just a few computation times required. Additionally, it has the lowest minimum description length (MDL) and the lowest average squared classification error (ASCE) and does it in the shortest amount of time. In another study, SVM [20] was used to diagnose breast cancer with a new version. Here, six types of SVM were described and utilized for evaluating performance. The findings of the standard SVM were compared to those of the other kinds. For both training and testing, four-fold cross-validation was used. St-SVM obtains 97.71% accuracy, 98.9% simplicity, and 97.08% sensitivity during training. In the testing phase, NSVM, LPSVM, SSVM, and LPSVM each achieved accuracy, sensitivity, and specificity of 96.5517%, 98.2456%, and 96.5517%, respectively.

To better identify breast cancer, the author's [21] employed inductive logic programming to classify the breast cancer data and offer an efficient technique. There was also a comparison study was conducted with a propositional classifier. As a performance metric, Kappa statistics, F-measure, ROC area under the curve, true-positive rate, and so on were computed. Two platforms, Aleph and WEKA, were used to mimic the system. In another study, jahjharia et al. [22] proposed decision tree algorithms for breast cancer diagnosis. The WEKA platform simulated the most popular decision tree algorithms, CART and C4.5, using MATLAB and python. The CART implemented in Python had the most fantastic accuracy (97.4%) and sensitivity (98.9%). The CART implemented in MATLAB had the highest specificity (95.3%), while the CART and C4.5 simulated in WEKA both had the lowest specificity (95.3%). The detail of followed dataset and experimentation is illustrated in the following section.

III.DATASET DETAILS

The breast cancer dataset was acquired from the University of California Irvine (UCI) machine learning repository [23]. There are 699 instances in this dataset, and the cases are classified as benign or malignant depending on their severity. Four hundred fifty-eight of these instances (65.50%) are mild, whereas two hundred forty-one (34.50%) are malignant. The class in the dataset is divided into two groups: two for the mild case and four for the malignant case, where two represent the soft case, and four represents the malignant case. The characteristics are included inside the dataset, which can be seen in Figure 1.



id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967

Figure 1: Characteristics of Dataset

IV.METHODOLOGY

To perform breast cancer detection, we performed several steps.

V. DATA PRE-PROCESSING

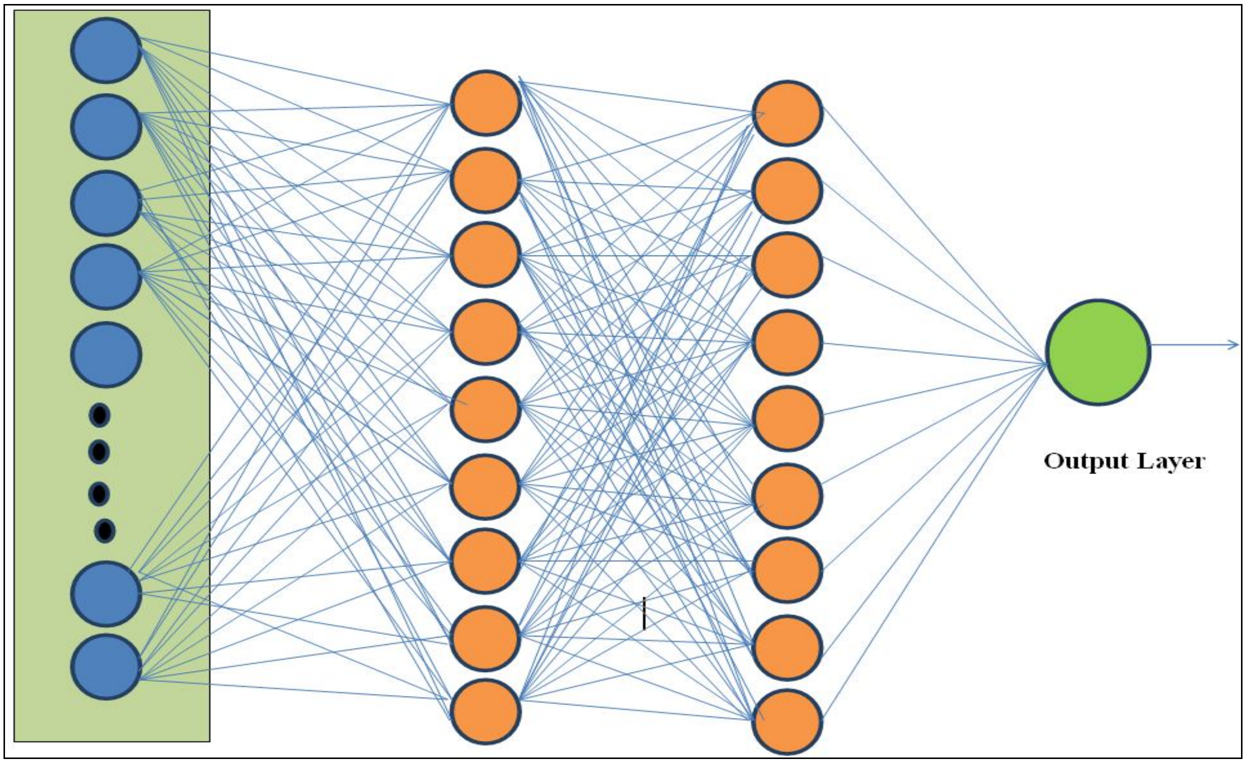
Data pre-processing is the first step in filling in the gaps left by missing data, detecting and eliminating outliers, and resolving self-contradiction problems. In the dataset, there are 16 missing values for characteristics that are not present. The mean takes the place of the missing attributes for that class. Additionally, the dataset is subjected to random selection to ensure that the data is adequately circulated. After data pre-processing, the dataset was divided into the training and testing phase. The training phase is used to extract the features from the dataset, and the testing phase is used to evaluate how the suitable model performs when it comes to predicting from the dataset. Each component of the dataset is split into two parts, Training and Testing. Cross-validation using K folds indicate that a single fold is used for testing, with the remaining K1 folds being used for training cyclically. Cross-validation is used to prevent over fitting in the data collection process. Specifically, a ten-fold cross-validation method is used to partition data, with nine-fold of the data utilized for training and one fold used for testing in each iteration of the research.

VI.CLASSIFICATION

Classification is the process of dividing a collection of data into categories. It may be done on both structured and unstructured data. Predicting the class of supplied data points is the first step in the procedure. Target labels and categories are all terms used to describe the course. Estimating the mapping function from the discrete input variables to the discrete output variable is classified as predictive modelling. The principal purpose is to figure out which category or class the new data belongs to. In this research study, we employed an Artificial Neural Network (ANNs) for breast cancer classification. The detail of the employed architecture is given below.

VI. ARTIFICIAL NEURAL NETWORKS (ANNS)

The artificial neural network algorithm is inspired by biological neurons and works by following the dendrite, soma, and axon workflow of biological neurons. Every ANN has an artificial neuron and a fundamental mathematical function as its internal structure. An artificial neural network's basic design consists of a collection of linked neurons organized into three layers: input, hidden, and output. This kind of network learns to execute tasks by taking into account a sufficient number of instances. The neural networks can be used for both classification and regression problems. The multilayer ANNs are the more advanced perception versions used to tackle complicated classification and regression problems. Perception is the essential kind of ANN used for binary classification. We also employed the ANNs for our classification task. The whole of neurons in the input layer of the ANN is equal to the number of characteristics in the dataset in its architecture. The hidden layer is another network component, with the number of hidden layers being counted as one layer. In this research, the input layer consists of 31 neurons that connect to 9 other neurons of the first hidden layer. There exist 9 -9 mapped connections between the first hidden layers to the second hidden layer. As the problem is a binary classification problem, there is just one neuron in the output layer. The employed architecture is illustrated below.



Input Layer
31 Neurons

Hidden Layer # 1
9 Neurons

Hidden Layer # 2
9 Neurons

VII. RESULT AND DISCUSSION

To determine if a cell is benign or malignant, we employed machine learning methods, i.e., Artificial Neural networks. We utilized a PC powered by an Intel Core i7 processor with 32 GB of RAM for processing reasons. The open-source machine learning package Scikit-learn written in the Python programming language is used. We also used Jupyter Notebook, an open-source online platform that allows us to create and distribute reports that contain live code, graphics, equations, and narrated text. The cross-validation was carried out using a ten-fold technique, which meant that the dataset was divided into ten equal groups. The deliberated model is validated using the ten-fold cross-validation method. The model was tuned for 100th epochs with five batch sizes, and the relu activation function is used in the hidden layers while the sigmoid is used at the output layer. The loss value is calculated using the Cross-Entropy loss function. The ANNs outperformed and computed 98.24% accuracy. The model training and accuracy graph is illustrated in Figure 3 and Figure 4, while Figure 5 presents the confusion matrix.

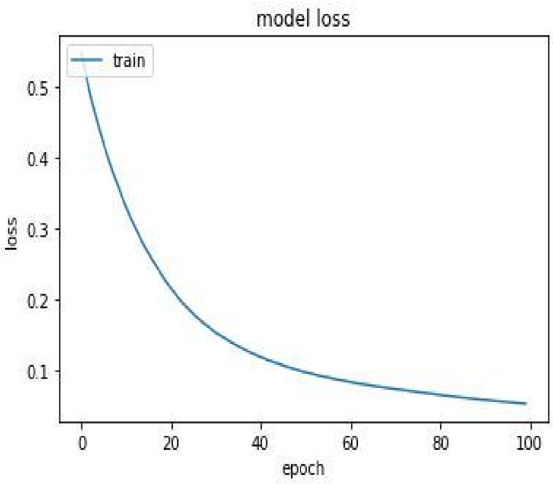


Figure 3: Model Training As a function of epochs

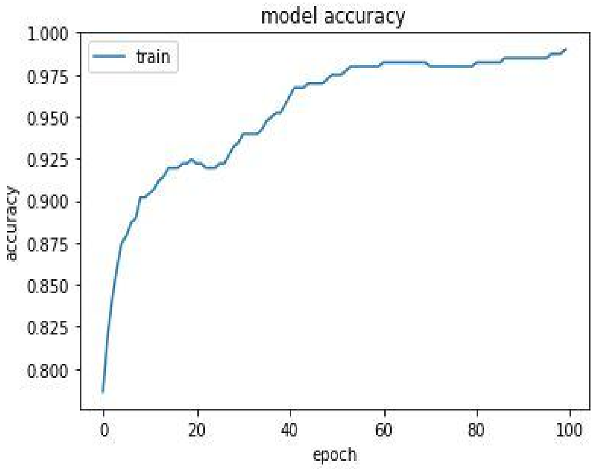


Figure 4: Classification accuracy as a function of epochs

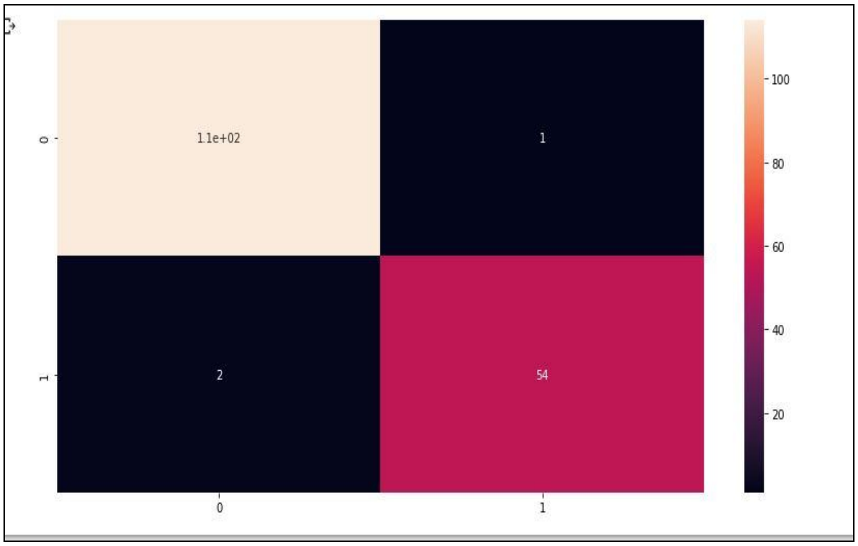


Figure 5: Confusion Matrix

VIII. CONCLUSIONS

This paper presented a machine learning technique for the prediction of breast cancer. It is incredibly costly and time-consuming to conduct a medical diagnostic process in the area of medicine. According to the system's recommendations, machine learning techniques may be used as a clinical assistant to detect breast cancer, which will be very beneficial for new doctors of a physician in the event of a misdiagnosis. The model produced by ANN is more consistent than any other method previously mentioned, and it has the potential to make essential advancements in breast cancer prediction. Based on the research findings, we can infer that machine learning techniques can automatically detect the disease with high accuracy.

REFERENCES

[1] Park SH, Han K. Methodological guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiol Soc N Am. 2018;286(3):800–9.

[2] Breast Cancer: Statistics, Approved by the Cancer.Net Editorial Board, 04/2017. [Online]. Available: http://www.cancer.net/cance r-types/breast-cancer/statistics. Accessed 26 Aug 2018.

[3] Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, Nakamura S. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-feld digital mammography in a population of women with dense breasts. Springer. 2016;24(1):104–10.

[4] Kurihara H, Shimizu C, Miyakita Y, Yoshida M, Hamada A, Kanayama Y, Tamura K. Molecular imaging using PET for breast cancer. Springer. 2015;23(1):24–32

[5] Azar AT, El-Said SA. Probabilistic neural network for breast cancer classifcation. Neural Comput Appl. 2013;23(6):1737–51.

[6] Nagashima T, Suzuki M, Yagata H, Hashimoto H, Shishikura T, Imanaka N, Miyazaki M. Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer. Springer. 2002;9(3):226–30

[7] Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS. Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Springer. 2013;22(2):153–60.

[8] Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020;. https://doi.org/10.1080/03772 063.2020.1713916.

[9] Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci. 2020;1(4):206.

[10] Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229

[11] Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK. Performance evaluation of random forests and artifcial neural networks for the classifcation of liver disorder. In: Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1–5.

[12] Ayon SI, Islam MM. Diabetes prediction: a deep learning approach. Int J Inf Eng Electron Bus (IJIEEB). 2019;11(2):21–7.

[13] Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, 2020. pp. 1–20.

[14] Hasan MK, Islam MM, Hashem MMA. Mathematical model development to detect breast cancer using multigene genetic programming. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 574–579, 2016.

[15] Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 2018;6:29637–47.

[16] Juneja K, Rana C. An improved weighted decision tree approach for breast cancer prediction. In: International Journal of Information Technology, 2018

[17] Yue W, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018;2(2):13

[18] Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 2012;23(7–8):2387–403.

[19] Senapati MR, Mohanty AK, Dash S, Dash PK. Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl. 2013;22(1):125–31

[20] Azar AT, El-Said SA. Performance analysis of support vector machines classifers in breast cancer mammography recognition. Neural Comput Appl. 2013;24(5):1163–77.

[21] Ferreira P, Dutra I, Salvini R, Burnside E. Interpretable models to predict Breast Cancer. In: Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1507–1511.

[22] Jhajharia S, Verma S, Kumar R. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. In: Proc. International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1–7.

[23] Breast Cancer Wisconsin (Original) Data Set, [Online]. https:// archive.ics.uci.edu/ml/machine-learning-databases/breast-cance r-wisconsin/breast-cancer-wisconsin.data. Accessed 25 Aug 2018