Proceedings of the Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)
IEEE Xplore Part Number: CFP21OSV-ART; ISBN: 978-1-6654-2642-8

Accurate Breast Cancer Detection and Classification by Machine Learning Approach

1st D. Sandeep
MTech Student
Computer Science and Engineering
GRIET, Hyderabad, Telangana, India

2nd Dr.G.N. Beena Bethel
Professor
Computer Science and Engineering
GRIET, Hyderabad, Telangana, India

***Abstract*— In this paper there is comparison of four different machine learning algorithms such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Fuzzy logic and Genetic algorithm on Wisconsin Breast Cancer Diagnosis (WBCD) dataset for the detection of breast cancer in women. The test accuracies are compared to show the efficient algorithm for the detection of breast cancer using those algorithms. The dataset is partitioned to 70% training data and 30% testing data. The results for the applied algorithms are CNN acquired 96.49% accuracy, RNN acquired 63.15% accuracy, fuzzy logic acquired 88.81% accuracy, and genetic algorithm acquired 80.399% accuracy.**

Keywords*— *Machine learning; Convolutional Neural Network; Recurrent Neural Network; Fuzzy logic; Genetic algorithm; Wisconsin breast cancer diagnosis (WBCD).

I. INTRODUCTION

Breast cancer is one of the dangerous diseases caused for women. The mortality rate is high if it is not treated in time. Breast cancer is caused due to the formation of tumors in the breast of women. There are two types of tumors that can be formed in body, they are benign tumor, malignant tumor. Benign tumors are harmless and they do not spread throughout the body they can be removed by proper medication, whereas malignant tumors are harmful they are cancerous tumors they can spread throughout the body if can't identified quickly. They may even cause death, so they need to be identified in the initial stage so that the harm can be reduced by proper medication or surgeries. According to World Health Organisation (WHO) in 2020 there were 685000 deaths, 2.3 million women diagnosed with breast cancer globally, there are 7.8 million women alive at the end of 2020 who are diagnosed for the past 5 years [7]. There are certain factors which causes breast cancer for women such as harmful use of alcohol, radiation exposure, family history, obesity, reproductive history. The breast cancer survival rates for at least 5 years after diagnosis ranges from 66% in India, 40% in South Africa.

For the diagnosis of the breast cancer the tumors must be detected, this can be done by ultrasound, mammogram, biopsy, magnetic resonance imaging (MRI). Machine Learning has multiple applications in medical field which can be used for the detection of various diseases. It allows a machine to learn by itself without human intervention, it can be used to train the models and test them. ML can be used in predicting the cancer cells, survival rates and specified treatment for patients. As mentioned in the Literature Review many authors have used different machine learning algorithms for breast cancer detection, but the main problem is to use the correct algorithm for efficient results. In this paper we use four different types of algorithms such as CNN, RNN, Genetic algorithm, and Fuzzy logic for detection of cancer tumors. The main goal of this paper to find out which machine learning algorithm is perfect for the tumor detection. As to know the perfect algorithm for the problem the above-mentioned algorithms results are compared.

II. LITERATURE REVIEW

There are many techniques proposed by many authors for the detection of breast cancer in women.

The authors in paper [1] proposed the Convolutional neural network on the mini-MIAS dataset which consists of greyscale mammogram images and obtained 82.7% accuracy. In paper [2] the authors proposed the CNN by using pre trained VGG-16 model on MIAS dataset which consists of mammogram images by using 2 main angles such as CC and MLO view and acquired the accuracy of 0.931 for CC view and 0.887 accuracy for MLO view.

Shaker K. Ali, Wamidh K. Mutlag proposed the fuzzy logic on the breast cancer dataset and obtained the accuracy of 98%, they divided their work into 2 parts one for identification and second for classification of cancer tissue [3]. In paper [4] authors proposed different data mining tools on WBCO dataset by using the classifiers such as Naïve Bayes, Bayesian Logistic Regression, simple CART, J48 and obtained the accuracy of 95.2% for Naïve Bayes, 65.42% for Bayesian Logistic Regression, 98.13% for simple CART, 97.2% for J48.

In paper [5] the authors proposed the genetic algorithm with phylogenetic trees and local binary patterns and support vector machines on the DDSM dataset and thus obtained the accuracy of 92.99% for genetic algorithm and 83.70% for local binary patterns and SVM.

The authors from paper [6] used support vector machines, random forest and Bayesian network on WBCO dataset and

acquired the accuracy of 97% for SVM, 96.6% for RF and 97.1% for Bayesian network.

In paper [9] authors used GRU-SVM, multilayer perceptron, linear regression, softmax regression, SVM, and nearest neighbours on WDBC dataset and acquired the accuracy of 90.68% for GRU-SVM, 96.92% for multilayer perceptron, 92.89% for linear regression, 97.36% for softmax regression, 97.7% accuracy for SVM.

The authors from [10] proposed genetic algorithm for breast cancers and different classifiers like J48, JRIP, Naïve Bayes on dataset from UCI which contains MSM-T and acquired the accuracy of 95.32% for J48, 92.98% for Naïve Bayes and 97.07% for JRIP.

In the paper [11] authors proposed ML algorithms like Logistic regression, SVM, KNN on WBCD dataset and by projected ensemble voting techniques then acquired the precision of 98.50%. From paper [12] authors used Fuzzy c-means algorithm along with pattern recognition model on WBCD dataset and acquired accuracy of 100% TP, 87% TN, 0% FP, and 13% FN.

From paper [13] authors used different data mining techniques such as SMO, IBK (KNN classifier), Best First trees on WBC dataset and acquired the accuracy of 95.46% for BF, 95.90% for IBK, and 96.19% for SMO. In paper [14] authors used different ML algorithms like DT, ANN, and SVM on ICBC from 1997 to 2008 datasets and acquired the accuracy of 0.936 for DT, 0.947 for ANN, and 0.957 for SVM.

In the paper [15] authors proposed ML algorithms like XGBoost, RF, and DNN on the information from BCIMS present at west china hospital of Sichuna University is used and acquired the results as 0.742 for XGBoost, 0.728 for RF, and 0.728 for DNN algorithm.

In related paper [16] authors proposed data mining techniques such as sequential minimal optimization, J48 on WBC dataset and acquired the accuracy of 96.99% for SMO, and 75.52% for J48.

The authors in the paper [17] proposed a supervised multinomial Bayesian learning for breast cancer detection using terahertz (THz) imaging. The work is done on the freshly excised murine tumors, the tumor is placed over filter paper and dried from excessive fluids and those are placed over scanning window for imaging process and histopathology process. Then the data is pre-processed and the proposed method is applied on them acquired the increase in the area under cancer and muscle ROC curve of 92.71%, 86.18%.

In paper [18] authors proposed ML algorithms such as Decision tree (J48), Naïve Bayes (NB), and Sequential minimal optimization (SMO) on the Wisconsin Breast Cancer and Breast cancer dataset. In this the 10-fold cross validation is used to assess the results, in breast cancer dataset J48 has highest accuracy of 98.20% and in WBC dataset SMO has the highest accuracy of 99.56%.

In related paper [19] authors proposed an optimized K-Nearest Neighbor model is used for the breast cancer detection. Wisconsin Breast cancer dataset is used on which the proposed model is applied. In this the hyper-parameter tuning on KNN is tested experimentally and acquired the performance of 90.10% and accuracy for proposed optimized KNN model is 94.35%.

From paper [20] proposed ML algorithms like SVM, KNN, Linear Regression on the Wisconsin Breast cancer dataset. A predicted system is proposed for early breast cancer detection by analysing smallest set of attributes from clinical dataset. Thus, the maximum classification accuracy obtained is 99.28%, and KNN has the highest accuracy than the linear regression and SVM.

In paper [21] authors considered a grey scale lung image and used malignant calculation through data and CPU infrastructure for improving the probability for finding lung cancer in the patient. The paper [22] is a survey on medical imaging of EIT by variable current pattern methods the authors have considered various papers and concluded that by electrical impedance tomography (EIT) is very useful for diagnosis sector of pulmonary problems in clinical decision making with high accuracy.

III. DATASET

Wisconsin Breast Cancer Diagnosis (WBCD) dataset is used in this work. Which is considered from UCI repository. This data is collected by Dr. William H. Wolberg from general surgery dept., W. Nick Street from computer science dept., Olvi L. Mangasarian from computer science dept. from university of Wisconsin [8].

3.1. Dataset Information

The dataset consists of the Fine Needle Aspirate (FNA) from a digitized image of the breast mass. It consists of 569 instances with 32 attributes, there are 10 real valued features which are computed for each nucleus of the cell such as a) radius, b) texture, c) perimeter, d) area, e) smoothness, f) compactness, g) concavity, h) concave points, i) symmetry and j) fractal dimension these features are computed on masses of breasts which can be divided to M= malignant and B= benign for diagnosis, this sample consist of 212 malignant tumors and 357 benign tumors.

IV. METHODS AND METHODOLOGY

The dataset may contain missing, uncertain data and noisy data values due to huge size and they are obtained from miscellaneous sources. So, the dataset is pre-processed to find any such type of values, pre-processing can be done manually by finding any such values and replacing them with mean of the values. The WBCD dataset does not contain any missing values so there is no need for replacement of the values. Google Colaboratory is used as simulation tool in the work for analysis.

a) Convolutional Neural Network

Convolutional neural network is a type of artificial neural network which can be used for many classification tasks like images or data values etc. It consists of input layer, output layer and hidden layers. In the input layer the data on which the operations to be done are given and then these data items go to the hidden layers.

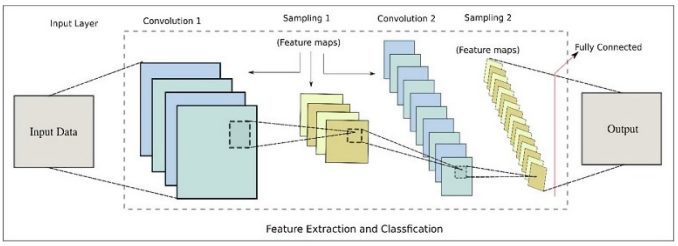

Fig. 1. CNN architecture

There may be more than one hidden layer in neural networks, the CNN has Convnet layer, pooling layer, activation layer these layers can be arranged in different ways along with their dense so that the operations can be done accurately and the data is transferred to output layer which gives the output for the performed operation.

In this paper conv1D layer, LeakyReLU layers are used in the hidden layers. Conv1D is used for 1-dimensional data in which kernel moves in one direction from input to output data can be used for sequences. The Conv1D is a convolution kernel that is convolved with layer input over a single spatial dimension to produce tensor of outputs, this layer is used for pattern recognition by extracting features from vectors. Those features are sent to next layers where LeakyReLU activation function used then the values is sent to the hidden layers, in proposed method two hidden layers are used and a sigmoid output layer is used.

LeakyReLU- Leaky Rectified Linear Unit is improved version for Rectified Linear Unit activation function. Both activation functions do not activate all the neurons at same time In ReLU if the function values go to negative input, then those values are returned as zero and for positive values it gives the value itself. Since the negative values are returned zero the neurons at those values are deactivated and causes dying ReLU, to overcome this dying ReLU LeakyReLU had been improved. In LeakyReLU modifies the function to allow small negative values instead of returning them to zero. Thus, the neurons at those values are active and we can overcome dying ReLU.

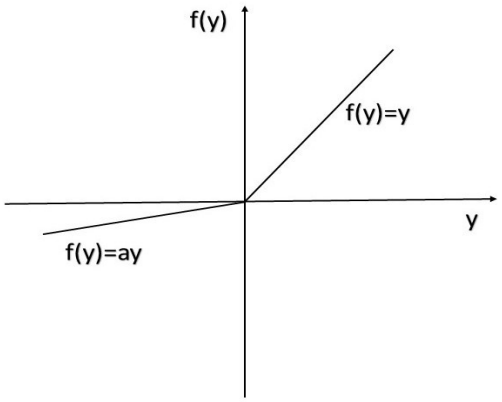

Fig. 2. LeakyReLU

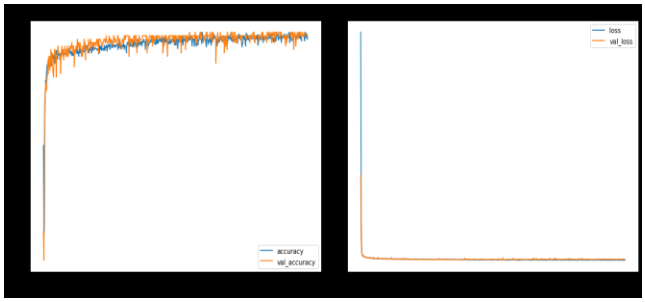

Fig. 3. Accuracy and loss values progress graph after applying CNN on WBCD dataset.

b) Recurrent Neural Network

Recurrent neural network is one of the artificial neural networks which can be done operations on data type or audio files etc. RNN helps to model a sequential data. It is similar to neural networks but a memory state is added to neurons. In neural networks all the input and output values are independent to each other but in RNN the hidden layers remember the data which is done in one previous layer and gives that remembered data to next layer as input so that it can perform operations on data. In the proposed method two hidden layers are used where the backpropagation is done among them by remembering the values from previous layers. The sequential model is used as input and a sigmoid function is used for the output from the neural network. It works similar to human brains for delivering predicting results.

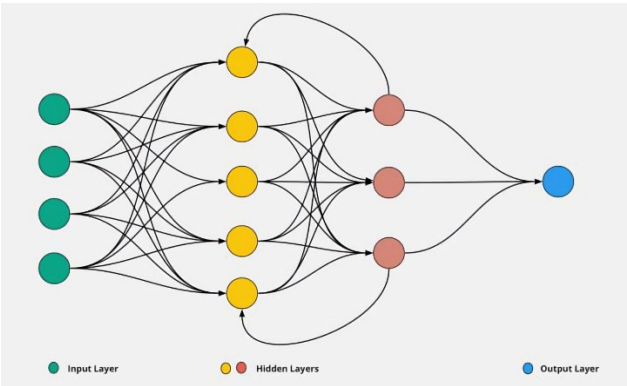

Fig. 4. Recurrent Neural Network architecture

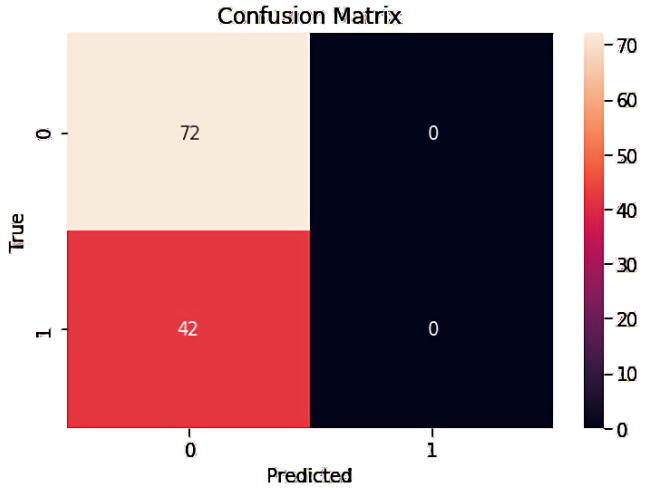

Fig. 5. Confusion Matrix after applying RNN to WBCD dataset

c) Genetic Algorithm

Genetic algorithm is a machine learning algorithm which uses a way of natural selection and genetics. It is a way to select in a natural way of species which can adapt and survive to the changes in environment. It is based on survival of fitness, each population consists of many individuals where each can be represented as int/bit/float/char value. These values are considered as a parent and they are used to obtain new offspring which is better than the considered data values. There can be mutation among them, adaption of genes, elitism, and selection of new population to get the better outcomes.

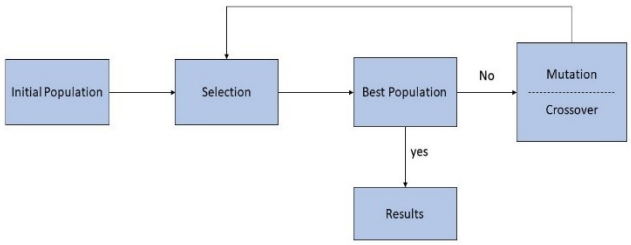

Fig. 6. Genetic algorithm

At first the initial population is considered and a sample is selected from that population and if results are good then those values are considered, if population sample results are not good then mutation and crossover is done to the population sample to acquire new population sample whose values are good.

In the proposed method the fitness of the model is calculated if the fitness is not good the mating is done along with mutation and random genome so that the acquired population fitness is better than the normal fitness so those population sample is used for acquiring the accuracy.

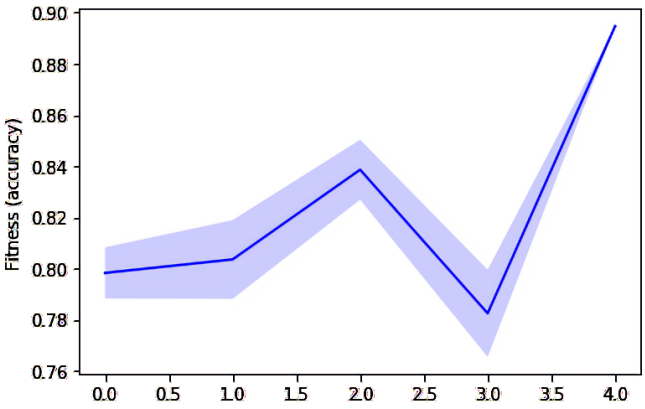

Fig. 7. Fitness accuracy graph applying genetic algorithm to WBCD dataset.

d) Fuzzy Logic

Fuzzy Logic is a machine learning algorithm which is similar to human reasoning, it makes a decision whether it may be yes or no such as Boolean output, so the fuzzy logic can be used for decision making and problem solving. The fuzzy logic consists of fuzzification module, knowledge base, inference engine, defuzzification module. The input is given to the algorithm then fuzzification module converts the input signals into different steps and transfers to knowledge base which contains the rules and inference engine which does human reasoning then the signals are transferred to defuzzification module which converts the received signals into human understandable form.

In the proposed method the input is taken from the dataset and those values are given to fuzzification layer which is used to machine readable format then those values are transferred to inference where the fuzzy rule is applied to those values and which can be used to detect the tumors and sent to defuzzification layer where the values are converted to normal outputs.

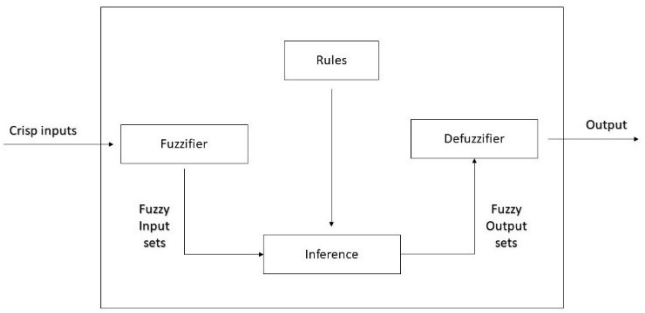

Fig. 8. Fuzzy Logic architecture.

4.1 Methodology

The Wisconsin Breast Cancer Diagnosis dataset is taken and then it is pre-processed for any missing values or noisy data, since the dataset is a clinical dataset, after importing the dataset it is checked for any missing values and those missing values are replaced by mean of the feature column, but there are no missing values then an encoder is applied so that the

objects are encoded into float values. Then that data is split into training and testing set and four different algorithms like Convolutional neural network, Recurrent neural network, Genetic algorithm and fuzzy logic are applied on it. After training them separately those algorithms are applied for testing phase and the accuracy of each algorithm is obtained and those accuracies are compared.

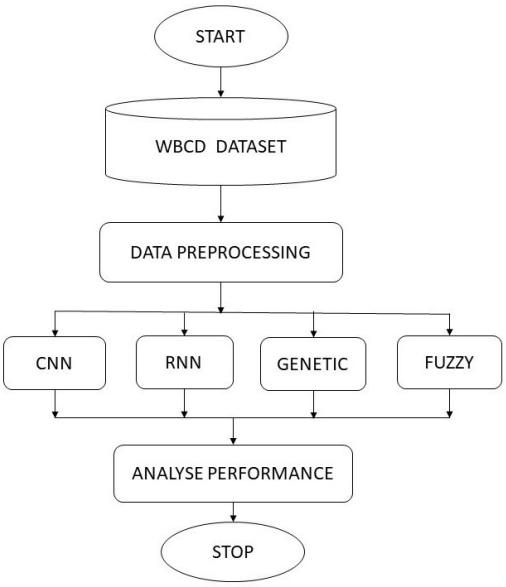

Fig. 9. Flow diagram of proposed breast cancer classification

V. RESULTS

The main concept of this study is to acquire a good accuracy for the detection of tumors in breasts of women. So, that different algorithms are used in this paper. The machine learning algorithms which are applied here can be useful for accurate detection of tumor cells. Then those algorithms are applied for the considered dataset to acquire the appropriate results. The CNN obtained the accuracy of 96.49%, RNN obtained accuracy of 63.15%, Genetic algorithm obtained accuracy of 80.39%, and Fuzzy Logic obtained the accuracy of 88.81%. CNN acquired highest accuracy for breast cancer identification on WBCD dataset.

TABLE 1

COMPARISON OF ACCURACY OBTAINED BY DIFFERENT ALGORITHMS USED FOR WBCD DATASET

ALGORITHMS USED	ACCURACY PERCENTAGE
Convolutional neural network	96.49
Recurrent neural network	63.15
Genetic Algorithm	80.399
Fuzzy Logic	88.81

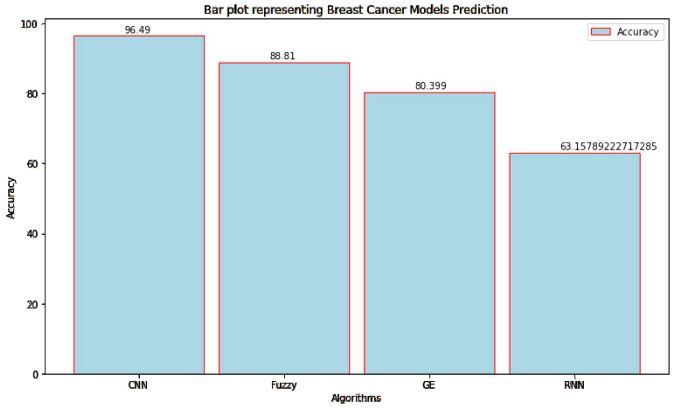

Fig. 10. Algorithm and Accuracy comparison

VI. CONCLUSION

Machine learning algorithms can be widely used for various medical field which can be used for good diagnostic tool for physicians to analyse the available data for further references. This paper represents the usage of four ML algorithms like CNN, RNN, Genetic algorithm, Fuzzy Logic and those accuracy is compared. CNN has acquired highest accuracy of 96.49%.

VII. FUTURE WORK

Although the proposed work has acquired good results the medical field need to have a 100% perfect results without any flaws. So, in near future there may be various advanced changes of technology or many new techniques may be available. Using those advanced techniques there may be improvements in the tumor detection and classification.

REFERENCES

[1] Y. J. Tan, K. S. Sim, and F. F. Ting, "Breast Cancer detection Using Convolutional Neural Networks for Mammogram Imaging System" International conference on Robotics Automation and Sciences (ICORAS),1-5,2017, ieeexplore.ieee.org

[2] Shuyue Guan and Murray Loew, "Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks" 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1-8,2017, ieeexplore.ieee.org

[3] Shaker K. Ali, Wamidh K. Mutlag, "Early Detection For Breast Cancer By Using Fuzzy Logic" Journal of Theoretical and Applied Information Technology, 15 September 2018, Vol.96. No 17 www.jatit.org

[4] Dr. S. N. Singh, Shivani Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis" 2018 4th International Conference on Computing Communication and Automation (ICCCA)

[5] Wener Borges de Sampaioa, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Marcelo Gattass, "Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM"2015,

http://dx.doi.org/10.1016/j.eswa.2015.07.046, www.elsevier.com/locate/eswa

[6] Dana Bazazeh and Raed Shubair, "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" 2016, https://www.researchgate.net/publication/310589496

[7] World Health Organization (WHO) https://www.who.int/health-topics/cancer#tab=tab_1

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Abien Fred M. Agarap, "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" 2019, ICMLSC

[10] Ahmed Abdullah Farid, Gamal Ibrahim Selim1, and Hatem A. Khater, "A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast Cancer Detection", 2020, www.preprints.org

[11] Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru, "Breast Cancer Detection using Machine Learning Way ", 2019, International Journal of Recent Technology and Engineering (IJRTE)

[12] Indira Muhic, "Fuzzy Analysis of Breast Cancer Disease using Fuzzy c-means and Pattern Recognition", March 2013, SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING

[13] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", January 2014, International Journal of Innovative Research in Computer and Communication Engineering

[14] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence",2013, J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124

[15] Can Hou1, MPH, DPhil; Xiaorong Zhong, DPhil, MD; Ping He, DPhil, MD; Bin Xu, MSc; Sha Diao, MSc, Fang Yi, MSc; Hong Zheng, DPhil, MD; Jiayuan Li, DPhil, "Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development", june 2020, JMIR MEDICAL INFORMATICS

[16] Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, and Gunter Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques", 2020, link.springer.com

[17] Tanny Chavez, Nagma Vohra, Keith Bailey, Magda El-Shenawee, Jingxian Wu, "Supervised Bayesian learning for breast cancer detection in terahertz imaging", 2021, Biomedical Signal Processing and Control 70 (2021) 102949, 28 june 2021, www.elsevier.com/locate/bspc

[18] Basker.N, Theetchenya.S, Vidyabharathi.D, Dhaynithi.J, Mohanraj.G, Marimuthu.M, Vidhya.G, "Breast Cancer Detection Using Machine Learning Algorithms", Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2551 – 2562, 05 may 2021, http://annalsofrscb.ro

[19] Tsehay Admassu Assegie, "An optimized K-Nearest Neighbor based breast cancer detection", Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020, http://journal.umy.ac.id/index.php/jrc

[20] Madhu Kumari, Vijendra Singh, "Breast Cancer Prediction system", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 2018, www.elsevier.com/locate/procedia

[21] Manoharan, Samuel. "Early diagnosis of Lung Cancer with Probability of Malignancy Calculation and Automatic Segmentation of Lung CT scan Images." Journal of Innovative Image Processing (JIIP) 2, no. 04 (2020): 175-186.

[22] Adam, Edriss Eisa Babikir. "Survey on Medical Imaging of Electrical Impedance Tomography (EIT) by Variable Current Pattern Methods." Journal of ISMAC 3, no. 02 (2021): 82-95.