International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8080
***Article Received:** 26 June 2023 **Revised:** 21 August 2023 **Accepted:** 01 September 2023*

Enhancing Breast Cancer Prediction through Deep Learning and Comparative Analysis of Gene Expression and DNA Methylation Data using Convolutional Neural Networks

Bandi Vivek[1], Mrs Tupili Sangeetha[2], Mustafa Nawaz S M[3], S. Lincy jemina[4], S.Nihal[5]
[1]Department of Computer Science and Engineering,
Panimalar Institute of Technology,
Chennai, India -600123
vivekbandi03@gmail.com
[2]Department of Information Technology,
Rajalakshmi Engineering college,
Chennai, India -602105
sangeethask09@gmail.com
[3]Department of Computer Science and Engineering,
Sri Sairam Institute of Technology,
Chennai, India -600 044
salmannawaz81@gmail.com
[4]Department of Computer Science and engineering,
Panimalar Engineering college,
Chennai, India -600123
lincypit@gmail.com
[5]Department of Artificial Intelligence and Data Science,
Panimalar Engineering college,
Chennai,India-600123
nihalkrishna431@gmail.com

Abstract— Recent advances in the production of statistics have resulted in an exponential increase in the number of facts, ushering in a whole new era dominated by very large facts. Conventional machine-learning algorithms are unable to handle the most recent aspects of huge data. This is a fact. In order to make an accurate prognosis of breast cancer, researchers employ and evaluate three distinct computer programmes called Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT). Within the context of huge statistics, we explore the question of how breast cancer may be predicted in this particular research. Gene expression and DNA methylation are both taken into consideration as part of the analysis (GE and DM, respectively). The purpose of the work that we are doing is to increase the capacity of the Deep Learning algorithms that are now being used for typing by applying each dataset individually and together. As a result of this decision, the platform of choice is MATLAB. In the process of breast cancer prediction, the Convolutional Neural Network (CNN) algorithm is used. Comparisons of GE, DM, and GE and DM are carried out with the help of this method. The results of the CNN algorithm are compared to those of the RF algorithm. According to findings of the experiments, the scaled system that was presented works better than the other classifiers. This is due to the fact that using the GE dataset; it acquired the best accuracy at the lowest cost.

Keywords- Deep Learning; Breast Cancer; Gene Expression; DNA methylation; Random Forest; Decision Tree.

I. INTRODUCTION

In the year 4200 B.C., the Egyptian people were diagnosed with breast cancer for the first time. Cancer of the breast is a condition that originates in the breast tissue. Symptoms of breast cancer may include a breast lump, a change in the contour of the breast, dimpling of the skin, fluid pouring from the nipple, a hastily-inverted nipple, or a patch of skin that is red or scaly. Breast cancer is the most common form of cancer in women. The machine learning approach was extended in order to create the deep learning technique. Deep learning algorithms function in numerous hidden layers, which brings them closer to human learning than machine learning algorithms do. As a consequence, deep learning algorithms provide more accurate results when used to predict breast cancer.

A number of years ago, the field of medicine started making widespread use of machine learning techniques, namely for breast cancer analysis and prediction. Data about breast cancer may be analyzed using a variety of approaches, including gene expression (GE) and DNA methylation (DM), for example. In this study, the information about the patient's genes comes from their DNA. The information contained in a cell's DNA may be passed along from one cell to another through ribonucleic acid (RNA).

II. RELATED WORKS:

A hybrid model was presented by Nan wu, Jason Phang, and others in [1]. The hybrid model is a blend of ResNet-based networks and BI-RADS networks. The structure of the model is rather simple. Both the BI-RADS network and the ResNet-based network contain 22 layers each, with the BI-RADS network working to increase the resilience and performance of breast cancer prediction. The ResNet-based network uses input data to change the resolution of the network's output. When compared to utilizing any of the two networks on its own, the hybrid model's performance is much superior. Nikhilanand Arya and Sriparna Saha proposed a stacked ensemble model in their paper [2], which is a two-stage model. The first stage of the model is a convolutional network, which is used to perform the feature extraction, and the second stage performs four algorithms one by one, including SVM, Random Forest, Naive Bayee, and Logistic Regression, which ultimately predicted the better accuracy. Ravi K. Samala and Heang ping Chan, the authors of paper [3], have proposed a digital Breast Temosynthesis (DBT) approach. This approach uses AlexNet CNN structure, and within this structure there are five layers: the first three layers are a fully connected layet, and the next two layers are a max pooling layer and a normalization layer. At long last, DBT provided a correct diagnosis of the breast cancer. The authors of the [4] Kihan Park and Wenjin Chen, along with their coauthors, said that the purpose of their work was to verify a novel biomarker for breast cancer and to suggest a design for a biochip diagnostic tool that they referred to as a MicroElectroMechanical System (MEMS). The breast tissue samples serve as the input to the system, and the system is ultimately able to identify between individuals with and without breast cancer. A framework for unsupervised feature learning was suggested in [5,] which was written by Dejun Zhang, Lu Zou, and others. During the phase of feature learning, a combination of principal component analysis and autoencoder method may be utilized to represent the feature in gene expression data. During the phase of classifier learning, the Adaboost algorithm will be used to predict breast cancer. Radial-based function neural network model (RBFNN) and ensemble boosting learning approach were both introduced by Ahmed Hamza Osman and Hani Mosetque abdullan Aljahdali, authors of paper number 6. The ensemble boosting approach is used to construct a series of processes in order to get a better prediction. In the end, the results are compared to the K-Mean algorithm and the Naive Bayes algorithm; however RBF-EBL gives a higher level of accuracy.

III. EXISITING SYSTEM

In the current system, in order to make a prediction about breast cancer, three distinct algorithms, namely Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) methods, are used. In current machine learning approaches, support vector machine is one of the algorithms that is considered to be the most popular. to provide really excellent classification results when working with data sets of a reasonable size. In the SVM classifier, the data samples are split up using the hyper plane. The hyper plane serves to differentiate between the characteristics of the DNA samples. The amount of computing work required is rather extensive. The Decision Tree algorithm will carry out a logical operation in order to make it simpler for us to choose a choice. The term "root" refers to the node that is at the very top of the tree. Every branch in the tree is responsible for carrying out its own logical action. The decision tree algorithm's conclusion is represented by the leaf node result as its final output. After the Random Forest method comes the ensemble classifier as the subsequent algorithm. It is made up of several decision trees that all branch out from the same root node. In the procedure for the random forest, take the results for all of the leaf nodes of the decision trees. In conclusion, carry out the entropy operation so that you may have a superior outcome.

IV. PROPOSED SYSYTEM

The goal of this study is to provide a method of deep learning that may be utilized for the prediction of breast cancer. The work that was suggested consisted of three major steps: first, the data are preprocessed by applying missing techniques by imputation and normalizing the data; second, the recommended work was carried out. Second, using the concept of mutual knowledge, they chose the key characteristics to focus on. In conclusion, we classified the characteristics by using an algorithm called convolutional neural network (CNN), which was utilized to predict breast cancer.

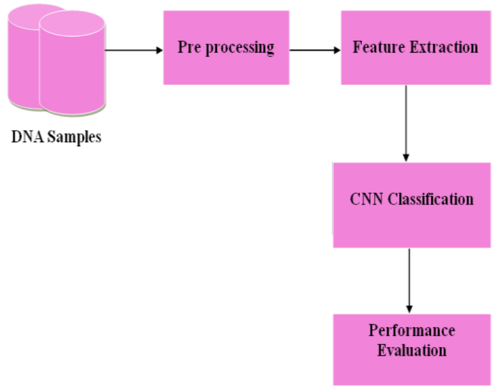


Fig 1: Architectural representation of proposed system

A. DATA COLLECTION

The DNA samples are being obtained via "The Cancer Genome Atlas –Data Portal," which may be found at Tega-data.nci.nih.gov. Last accessed on January 30, 2018, and may be found at https://portal.gdc.cancer.gov/. The DNA samples include two distinct types of data, referred to respectively as GE data and DM data. Each data sample consists of 254 different people, of whom there are two distinct groups: 215 patients (diagnosed with breast cancer), and 39 healthy individuals. The values of 16,077 genes are included in each sample.

B. DATA PREPROCESSING

In the process of data preparation, the purpose of this module is to transform the DNA samples into a structure (table) format, with each row belonging to a patient and each column relating to a set of gene data. Use of the Least Mean Square (LMS) algorithm, which is one sort of adaptive filter, is included in this section. This filter will clean the DNA samples by removing all of the noisy data. The LMS algorithm's purpose is to reduce the computational complexity as much as possible.

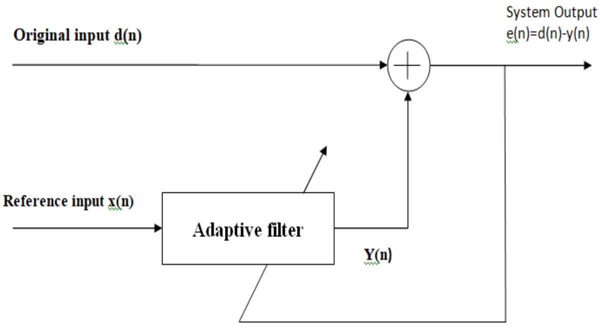


Figure2. Block Diagram of LMS Algorithm

C. FEATURE EXTRACTION

The quantity of resources that are required to characterize a large number of DNA samples may be reduced via the process of feature extraction. It begins with an initial set of measured data and constructs derived values (features) with the intention of being informative and non-redundant. This helps to facilitate the following learning and generalization phases, and in certain situations, it leads to improved human interpretations.

D. CONVOLUTIONAL NEURAL NETWORK

One of the types of algorithms used in deep learning methods is known as the convolutional neural network. The CNN algorithm has a greater number of convolutional layers, each of which may be coupled or pooled. It utilizes one of many different forms of multi-layer perception. More applications, including image resolution, video recognition, natural language processing, and image classification, are used using the CNN algorithm. Today, CNN is utilized in agriculture, and weather forecasts may be provided by Satellites LSAT. These Satellites can also anticipate the rise of produce. Pooling, fully connected, padding, and stride are some of the capabilities that are included in the CNN algorithm. Adjusting the dimension on each and every layer of the CNN might be considered pooling. When we talk about a network being completely linked, we imply that each layer of the network may connect to each other. The many layers of padding each have their own unique weight. Based on the stride value, the input value might cause a move from one hidden layer to another hidden layer. Stride is a performance that is performed on the value.

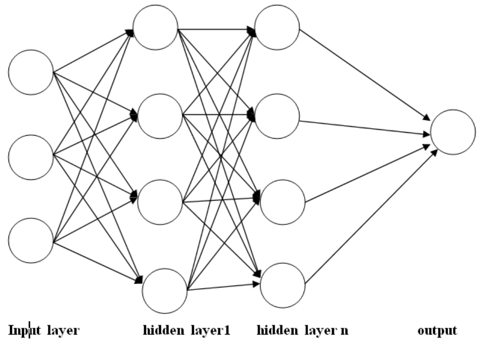


Figure3. Design of CNN Network

There are a total of four layers in the CNN design, and they are referred to as the CONV layer, the Relu layer, the pooling layer, and the fully connected layer. In order to extract the features of the DNA samples, the CONV layer is used. Radius, Texture, and Perimeter are the characteristics that have been retrieved. An activation function is applied in a feature-wise manner by the Relu layer. The sigmoid activation function is what we employ. Assuming that the output of the Relu layer contains negative values, the sigmoid function would then be engaged at the appropriate moment to convert the negative value into zero. The DNA samples are not altered in any way by this layer. The sigmoid function is used here.

$$\partial(X) = (1 + e^{-X})^{-1} \tag{1}$$

The next layer will act as a pooling layer, and its purpose will be to cut down on the spatial dimension of the input depending on the stride value. The max pooling filter is used here in this layer. When compared to other pooling filters, this filter produces the greatest results.

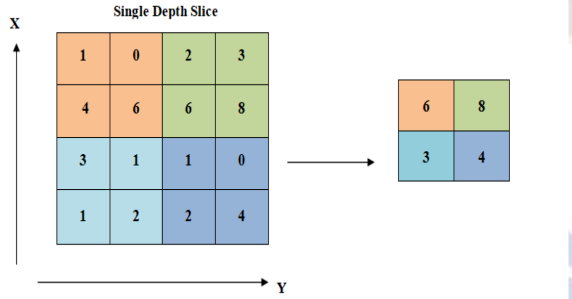


Figure4. Example of Max Pooling Filter

The fully connected layer of the convolutional neural network is the last layer. Every neuron in one layer of a network is connected to every neuron in all other layers of the network through this layer in the network.

V. RESULT AND ANALYSIS:

During this stage, the performance may be judged according to how accurate it was. In order to determine whether method, Random Forest or Convolutional Neural Network, is more accurate in predicting breast cancer, we will compare the two methods.

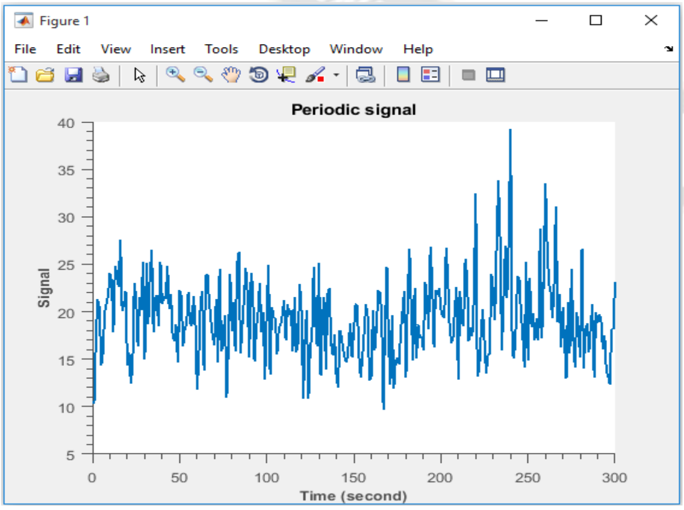


Figure5. Plot DM Sample with Noise Data

Shown in picture 5 is a representation of the DM samples plot using noise data.

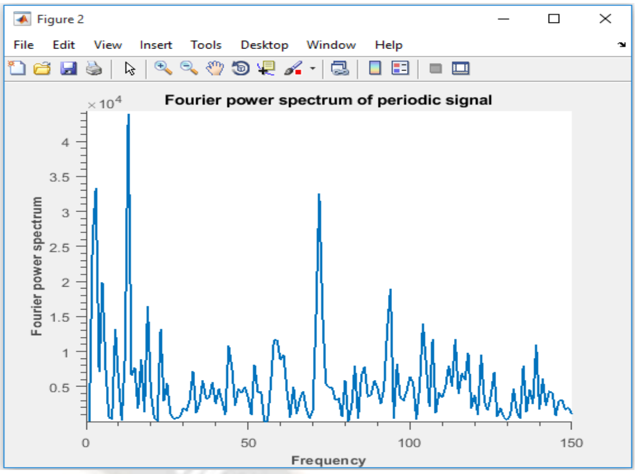
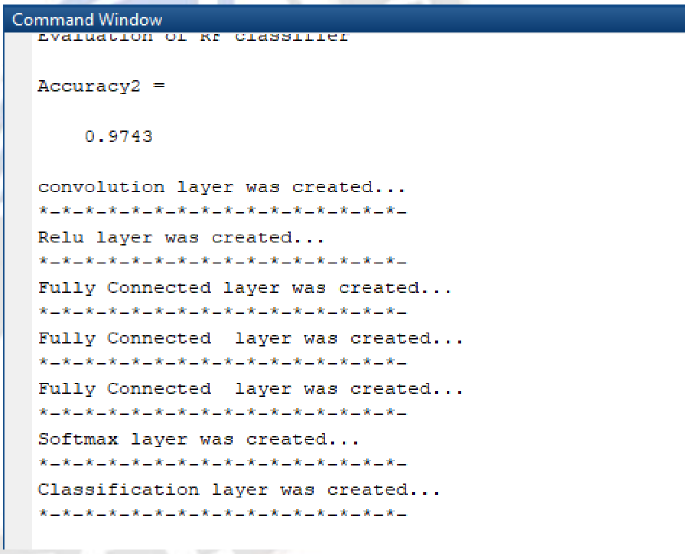


Figure6. Plot DM Sample without Noise Data

Figure 6 shows a representation of the DM sample that does not include noise data. To clean up the DM samples using the LMS adaptive filter and get rid of any noisy data.



```
Evaluation of RF classifier

Accuracy2 =

    0.9743

convolution layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Relu layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Fully Connected layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Fully Connected  layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Fully Connected  layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Softmax layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
Classification layer was created...
*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_
```

Figure7. Convolution Layer Created

All of the layers of the Convolutional Neural Network, which are represented by the number 7 in Figure 7, have been successfully formed. These layers include the CONV layer, the Relu layer, the pooling layer, and the Fully connected layer.

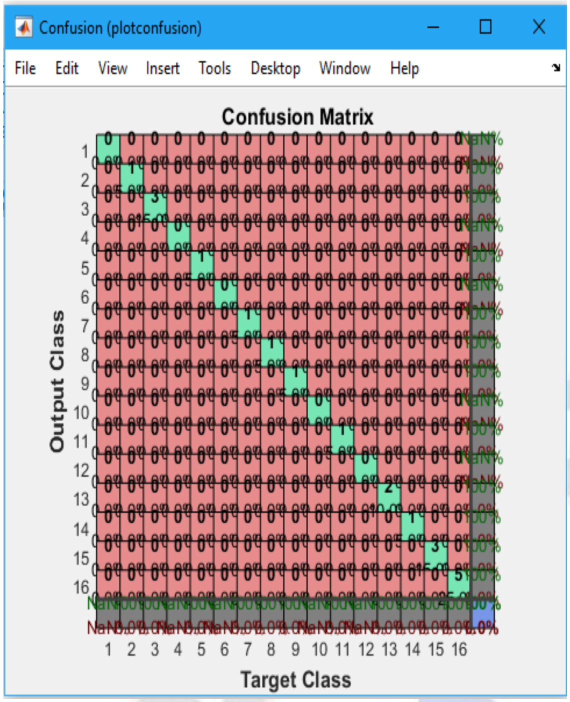


Figure8. Confusion Matrix

The Confusion Matrix is represented by the number 8 in the illustration. In the confusion matrix, recall, precision, and accuracy are measured based on true positives, true negatives, and false positives and false negatives, respectively.

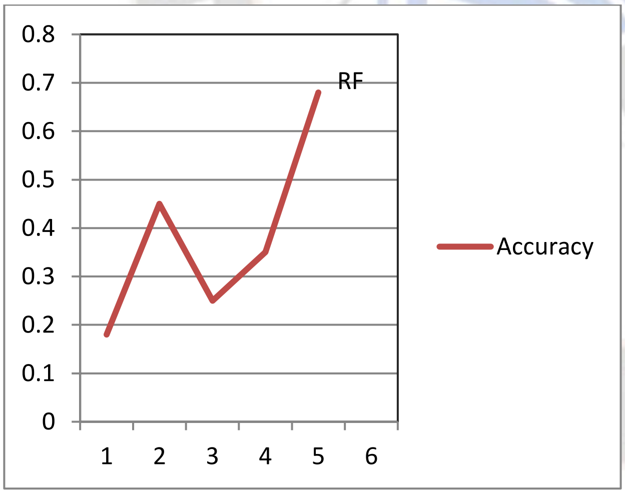


Figure9. Accuracy for RF

Figure 9 displays an illustration of how accurate the Random Forest algorithm is. The accuracy of each tree is shown on the graph, and the Random Forest technique is then used to execute an entropy operation. The accuracy of the RF algorithm was determined to be 68% once all was said and done.

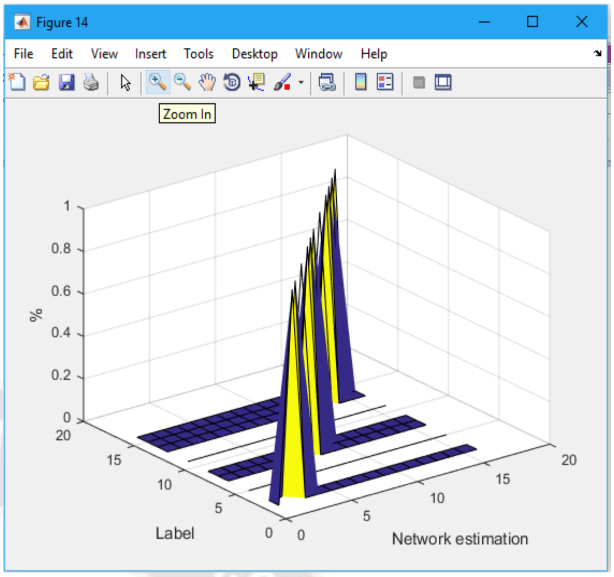


Figure10. Comparison of Accuracy in RF and CNN

Shown in figure 10 is the level of accuracy achieved by the Random Forest algorithm and the Convolutional Neural Network. Both the Random Forest method (68%) and the Convolutional Neural Network (97%) have a high degree of accuracy. The picture above depicts the RF with a yellow label, whereas the CNN is shown with a blue label.

VI. CONCLUSION:

The risk of developing breast cancer may be estimated using the deep learning technology. The objective of this project is to develop a medical data classifier that is both accurate and efficient computationally. In this study, our goal was to use CNN-based classification algorithms to examine a large data collection of GE breast cancer cases in order to make an accurate prediction of the cancer incidence. As the signaling platform, MATLAB served as the large data system that we utilized. The distinctive feature of our research approach is that we analyzed not one but two distinct kinds of big data, namely DM and a composite dataset that included both GE and DM. This was done in order to investigate the potential advantages of using both forms of data in the categorization of breast cancer. In the not too distant future, we are going to put this initiative into action employing a methodical optimization approach. This approach consists of identifying the most useful characteristics for the purpose of predicting breast cancer.

ACKNOWLEDGMENT

No funding sources.

AUTHOR CONTRIBUTION

Author 1 implemented the concept specified by the author 2 under the supervision of authors 3 & 4. The authors 3 & 4 & 5 drafted the article under the guidance of author 2.

DOI: https://doi.org/10.17762/ijritcc.v11i11s.8080
***Article Received:** 26 June 2023 **Revised:** 21 August 2023 **Accepted:** 01 September 2023*

REFERENCES

[1] A. H. Osman and H. M. A. Aljahdali, "An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model," in IEEE Access, vol. 8, pp. 39165-39174, 2020, doi: 10.1109/ACCESS.2020.2976149.

[2] D. Sun, M. Wang and A. Li, "A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, pp. 841-850, 1 May-June 2019, doi: 10.1109/TCBB.2018.2806438.

[3] Pattnaik, M. ., Sunil Kumar, M. ., Selvakanmani, S. ., Kudale, K. M. ., M., K. ., & Girimurugan, B. . (2023). Nature-Inspired Optimisation-Based Regression Based Regression to Study the Scope of Professional Growth in Small and Medium Enterprises. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 100–108. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2576

[4] D. Zhang, L. Zou, X. Zhou and F. He, "Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer," in IEEE Access, vol. 6, pp. 28936-28944, 2018, doi: 10.1109/ACCESS.2018.2837654.

[5] K. Park, W. Chen, M. A. Chekmareva, D. J. Foran and J. P. Desai, "Electromechanical Coupling Factor of Breast Tissue as a Biomarker for Breast Cancer," in IEEE Transactions on Biomedical Engineering, vol. 65, no. 1, pp. 96-103, Jan. 2018, doi: 10.1109/TBME.2017.2695103.

[6] Mr. Rahul Sharma. (2015). Recognition of Anthracnose Injuries on Apple Surfaces using YOLOV 3-Dense. International Journal of New Practices in Management and Engineering, 4(02), 08 - 14. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/36

[7] N. Wu et al., "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," in IEEE Transactions on Medical Imaging, vol. 39, no. 4, pp. 1184-1194, April 2020, doi: 10.1109/TMI.2019.2945514.

[8] Abdul Rahman, Artificial Intelligence in Drug Discovery and Personalized Medicine , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[9] R. K. Samala, H. -P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter and K. H. Cha, "Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning Using Deep Neural Nets," in IEEE Transactions on Medical Imaging, vol. 38, no. 3, pp. 686-696, March 2019, doi: 10.1109/TMI.2018.2870343.