Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Integrated Breast Cancer Analyzer and Predictor Using Machine Learning and Deep Learning

Divij Chawla[1], M. Pushpalatha[2], S. Poornima[3], Pragya Saxena[4]

[1]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur Campus, Chennai, India. E-mail: dr6136@srmist.edu.in

[2]Professor, Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur Campus, Chennai, India. E-mail: pushpalm@srmist.edu.in

[3]Assistant Professor, Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur Campus, Chennai, India. E-mail: poornims@srmist.edu.in

[4]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur Campus, Chennai, India. E-mail: ps8307@srmist.edu.in

ABSTRACT

Cancer, undeniably, is one of the most dangerous threats to public health worldwide. Breast Cancer in specific accounts for about 12.3% of all cancer cases, being the second most common type of cancer among women. Recent statistics reveal that Breast Cancer cases among Indian women are on the rise. With a new case coming up every 4 minutes, it's now prominently present even among younger age groups. Breast Cancer Detection requires a histopathologist to carry out a tissue biopsy & determine the presence of abnormal cells. Even the slightest slip in precision can result in False Positive or False Negative results. Over the years, medical advancements have made this process simpler. Nowadays, Machine Learning approaches such as K-Nearest Neighbour, Support Vector Machine, etc., are employed. These house effective diagnostic capabilities but aren't 100% accurate. Further, these achieve high accuracy only for Binary classification (Benign/Malignant) and support low accuracies for Multi-Class classification & early-stage detection, posing a need to improve diagnosis quality. In this paper, we represent the application of DenseNet Classifier and Random Forest Algorithm for Breast Cancer Detection, Classification & Prediction. We propose a robust integrated system capable of detecting Breast Cancer along its specific stage. Further, if a suspected patient comes out to be healthy, it will be able to; predict their chances of developing Breast Cancer in future. Results of this implementation show that the Random Forest Algorithm gives high accuracy of 97.07% for prediction. Similarly, the DenseNet Classifier achieves an exceptional accuracy of 97.59% for detection.

KEYWORDS

Breast Cancer, DenseNet, Invasive Ductal Carcinoma (IDC), MongoDB Atlas, Random Forest, Wisconsin Breast Cancer (Diagnostic).

Introduction

Several factors have been linked to Breast Cancer; these include family history, lifestyle factors, hormones and hormone medicines, the density of breasts, radiation etc. Breast Cancer develops when cells begin to grow out of control; cells usually develop a tumour that can often be detected with an X-ray or felt as a lump [1]. The cancer cells can develop in the lobules, the breasts' ducts, the breasts' fatty tissues, or fibrous tissues. Lobules [2] are the milk-producing glands, and ducts are the pathways through which milk flows to the nipples. Malignant tumours expand to the neighbouring cells, leading to metastasize or reaching other parts, whereas Benign masses can't grow to other tissues; the expansion is only limited to the Benign mass [3]. Due to the small size of the tumour at the early stages, there may be an absence of symptoms, and therefore detection of Breast Cancer is hard in the beginning. However, some signs of Breast Cancer can be a sudden change in the size or shape of the breasts, breast pain, bloody or unexplained discharge from nipples, changes in appearance or texture of the skin of the nipple or the whole breast might also be a symptom of Breast Cancer.

An accurate diagnosis can be defined as one that can differentiate between Malignant and Benign tissues and result in low False Positive and False Negative [4] cases. Machine Learning techniques have been popularly used in devising and evaluating algorithms for facilitating the classification and prediction of Breast Cancer. Methods like K-Nearest Neighbour, Naive Bayes Classifier, Decision Trees, Support Vector Machine, Artificial Neural Network etc., have contributed significantly to research work to create an effective system to detect Breast Cancer or estimate the probability of a person developing it in their lifetime.

In this paper, we analyze two algorithms, namely DenseNet Classifier, for detection and classification of Breast

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Cancer and Random Forest Algorithm, to help predict a person's probability for developing it. We use the Wisconsin Breast Cancer (Diagnostic) dataset for the Prediction module, and an image dataset consisting of Histopathological (IDC) images associated with 735 clinical cases of Breast Cancer. We determine the performance of the algorithms in terms of accuracy, training and testing process.

Related Works

Over the years, a significant amount of research and technological advancements have taken place in medical science. The primary motivation behind these has always focused on improving upon the currently existing mechanisms and devising new & efficient methods to help timely detect and cure Breast Cancer. A majority of these are mentioned below:

In 2020, Tanishk Thomas, Nitesh Pradhan and Vijaypal Singh Dhaka [5]; carried out a comparative analysis of numerous Machine Learning techniques (ML) such as Support Vector Machine, K-Nearest Neighbour, Naive Bayes, Decision Trees, etc. They employed these techniques to detect the presence of Breast Cancer at early stages. The main ideology behind this research work was to identify which Machine Learning Algorithm provides the best accuracy.

In 2019, Quang H. Nguyen et al. [6] employed Machine Learning approaches such as Voting Classifier, SVM, etc., on the Wisconsin Breast Cancer (Diagnostic) dataset, consisting of 570 raw data values. They further leveraged the Cross-Validation technique to divide the dataset into a 90-10 Train-Test ratio. The limitation of this approach was the requirement for odd-numbered base models. Further, it employs SVM, SGD, etc., models, which are considered to be Black-box models and hence aren't readily accepted.

In 2019, H. Zhang et al. [7] leveraged Convolution Neural Networks, along with a SE-ResNet-based module, to determine Breast Cancer's presence by analysing the Histopathological images. The main motivation behind this research work was to develop/invent a new & unique Neural Network.

In 2018, Kun Zhang et al. [8] made use of the Computer-Aided diagnosis and Pattern Recognition based techniques to predict the chances of developing Breast Cancer in the coming years. They employed this technique for carrying out Binary and Multi-Class classification. The limitation of this method is the challenging process of setting up a Convolutional Neural Network by fixing the number of layers. Even the slightest error in fixing the number of layers can adversely affect the performance.

In 2018, S. Pal et al. [9] employed the Data Mining approach to not only predict the chances of developing Breast Cancer but also determine if the cancer is mild or deadly.

In 2017, K. Srikanta Murthy and R. Sangeeth [10] employed Digital Image Processing techniques, incorporating techniques such as Morphological Operator, Otsu Algorithm, Sobel Edge Detection, etc., to segment the microcalcification cells for Breast Cancer Detection by classifying them into Benign and Malignant.

In 2017, K. S. Sim et al.[11] leveraged Convolution Neural Networks to capture and analyze the Breast Cancer tissue image. The main motivation behind carrying out this research was to boost the rate of early detection and effective treatment.

In 2016, Ladislav Lenc and Pavel Král [12] employed the technique of using Local Binary Patterns (LBP) to detect the presence of cancer among patients. They further employ the SVM Classifier and leverage the ROC curves as their performance measuring indices to determine the quality of the classifier built. They applied these techniques to the DDSM and MIAS datasets. One limitation of this approach was its requirement to set the LBP parameter correctly. In case there's even a minimal error in the LBP parameter value, it can lead to wrong Cancer classifications.

In 2013, Mahersia Hela et al. [13] made use of the Mamographical analysis technique, an X-ray mechanism, to determine the presence of abnormalities in the breast tissues. The major flaw in this approach is that it didn't provide a definitive diagnosis. Their approach could only detect the presence of abnormalities. To confirm the presence of Breast Cancer, further tests such as an MRI of the breast were necessary.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

In 2012, Daniel R.Bauer et al. [14] leveraged the concept of Microwave imaging to detect the presence of Breast Cancer. In their approach, they passed Microwaves into the breast tissue, converted them into Thermoacoustic waves and produced a dielectric loss. Based on the variation of these dielectric loss values across different parts of the tissue, their approach would detect the presence of cancerous tissues.

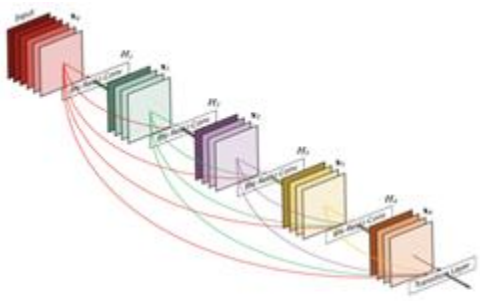
Background Knowledge

A. Deep Learning and Machine Learning Techniques

1) DenseNet Classifier

DenseNet (Dense Convolutional Neural Network) is a robust object-detection algorithm that concatenates the previous layers' information with the upcoming layers. Hence, the Feature Maps of each layer act as an input for the subsequent layers. It achieves this by working in contrast to the Traditional CNN [15], in which the number of layers are directly proportional to the connections:

'L' layers = 'L' Direct Connections



DenseNet Structure

$$a^{[l]} = g\left([a^{[0]}, a^{[1]}, a^{[2]}, \dots\dots\dots, a^{[l-1]}\right)$$

Figure 1. Visual Representation of the Dense Convolutional Neural Network [27] Structure

In DenseNet, each network hasL(L+1)/2 direct connections [16], where 'L' = Number of Network Layers**.**DenseNet comes in various types such as DenseNet121, DenseNet169, DenseNet201, etc. Here, in our project, we will be using the DenseNet201 Algorithm, where 201 represents the number of network layers.

DenseNet201 : 5+(6+12+48+32)*2 = 201

- 5 = Convolution & Pooling Layer
- 3 = Transition Layers (6,12,48)
- 1 = Classification Layer (32)
- 2 = Dense Block (1x1 and 3x3 Conv)

Our model leverages DenseNet201 for training purposes with a batch size of 16, input shape tuple as (50,50,3) and the weights as ImageNet.

DenseNet concatenates the result of one layer with another by leveraging a Composite Function consisting of the Convolution & Pooling Layers, Non-activation Layers and Batch Normalisation. It divides the entire network into different DenseBlocks, each having a different number of "Fitters" but the exact dimensions. It leverages these Transition Layers to apply processes [17] such as "Downsampling", "Batch Normalisation", etc. to ensure that all the Feature Maps are of identical dimensions for effectual training.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

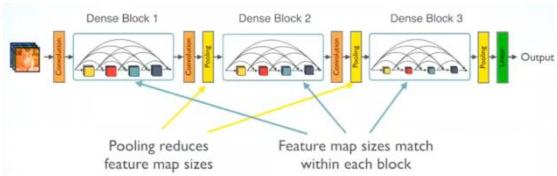


Figure 2. Dense Blocks [17] Structure in DenseNet

The growth factor "K" in DenseNet [17,18] represents the amount of information that travels from one layer to another**.**

$$K(l) = (K(0) + k(l-1)) \quad (1)$$

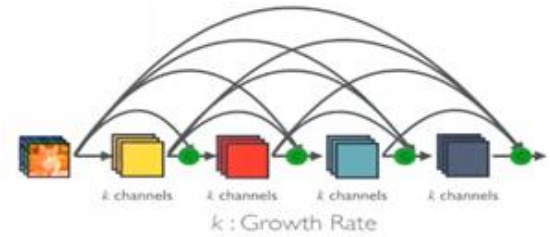


Figure 3. Visualization of Information Flow [17] between Channels based on 'K' Growth Rate

2) Random Forest Algorithm

Random Forest is a robust algorithm that leverages numerous Decision Trees in the form of an ensemble. It follows the ensemble technique and takes into account the prediction from all the Decision Trees by pooling them and predicting the final output.

It ensures a low correlation value [19] by considering a large set of uncorrelated trees together, making them operate as a single group/committee and thereby provides an accurate prediction by considering the feature having the maximum votes.

It further leverages the concept of Bagging, which helps ensure that each tree takes up a random sample of values from the entire dataset, thereby resulting in unique Decision Trees and accurate predictions. Random Forest further combines the power of Bagging with "Feature Randomness" [19,20] that allows each tree to take up only one specific feature out of many subsets. This works in contrast to Decision Trees, which take up all possible features and then choose a particular one.

In our project, we use the Random Forest Algorithm to train the medical dataset consisting of 33 attributes; we drop two columns (unnamed and id) as they are not relevant to the prediction process. We further perform exploratory data analysis on the dataset to remove none/null values, scale the parameters into the range of 0-1 and understand the feature importance.

Random Forest helps us find the importance of each feature by determining the Gini Importance value. The Gini Importance Value [21] is the probability of reaching the node and how the node impurity decreases.

$$ni_j = w_{left(j)}C_{left(j)} - w_{right(j)}C_{right(j)} \quad (2)$$

Equation (2) is the Mathematical Formula for Gini Importance Value.

3) Cross-Validation

Cross-Validation is a robust technique that helps statistically-segment/partition the dataset into the learning and testing sets that help train and test the model, respectively. It is one of the most efficient techniques that help streamline the evaluation process of the learning algorithm.
Cross-Validation randomly divides that dataset into testing and training partitions [22] such as 70% Train and 30%

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Test or 60% Train and 40% Test, etc. One of the most basic forms of Cross-Validation is the K-Fold Cross Validation technique that leverages one of the K-segments for the validation process. Some complicated types of Cross-Validation keep the base as the K-Folds.

Proposed Methodology

A. Datasets in Use

To power the knowledge engine for our “Integrated Breast Cancer Analyzer and Predictor" project, we are leveraging the best in class Image and CSV datasets. The CSV dataset is the Wisconsin Breast Cancer (Diagnostic) dataset [23] that supports our Prediction module and consists of 569 clinical cases.

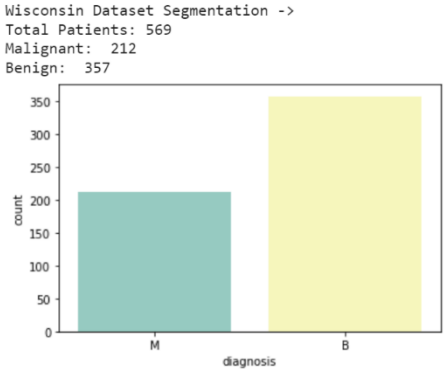


Figure 4. Graphical Description of Wisconsin Breast Cancer Dataset

Table 1. Description of Wisconsin Breast Camcer Dataset

Dataset Name	No. of Patients	No. of Classes	Malignant	Benign
Wisconsin Breast Cancer (Diagnostic)	569	2	212	357

The Image dataset consists of the Histopathological (IDC) images associated with 735 clinical cases of Breast Cancer. This dataset powers our Detection module and focuses on Invasive Ductal Carcinoma (IDC) [24], one of the most commonly found Breast Cancer subtypes among women.

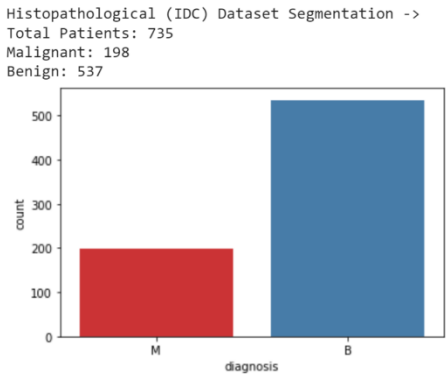


Figure 5. Graphical Description of Histopathological (IDC) Dataset

Table 2. Description of Histopathological (IDC) Dataset

Dataset Name	No. of Patients	No. of Classes	Malignant	Benign
Histopathological Invasive Ductal Carcinoma (IDC)	735	2	198	537

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

B. Architecture

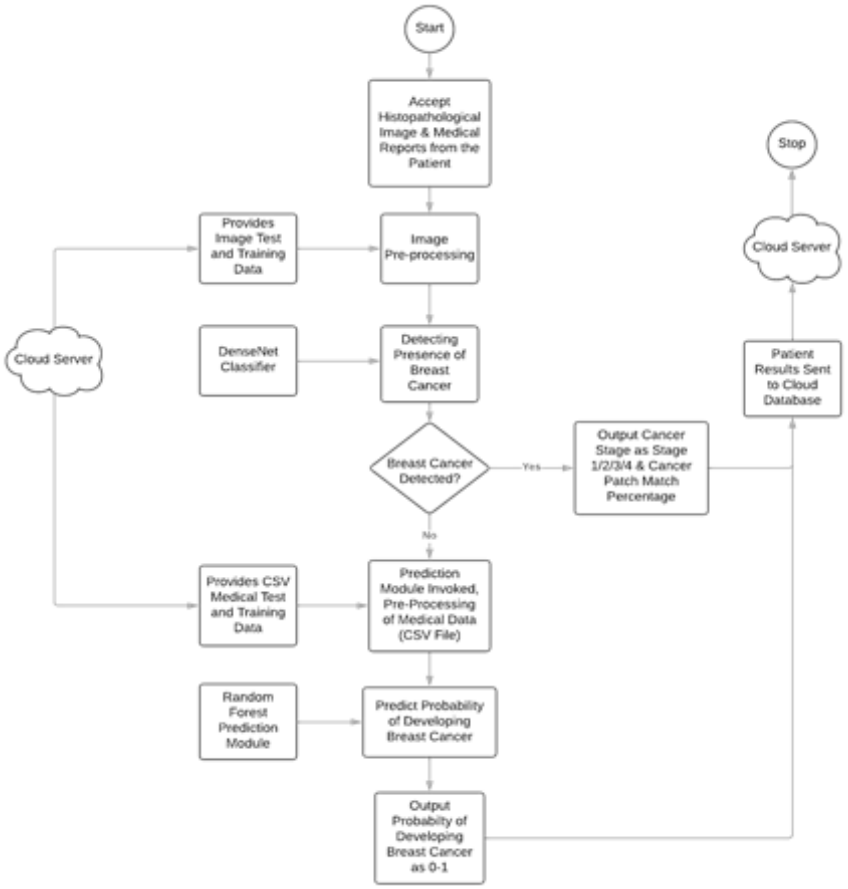


Figure 6. Proposed Integrated Breast Cancer Detection and Prediction Module

Our "Integrated Breast Cancer Analyzer and Predictor" application houses an easy to use Python Tkinter based user-interface that allows patients to provide their details such as their name, age, symptoms, etc. It will further accept the Histopathological image from the patient.

Once the patient has provided the image, it will send the patient details to our MongoDB Atlas (AWS Instance) based Cloud storage. It will further invoke our DenseNet based classifier that will help analyze the Histopathological image and detect the presence of Breast Cancer. If Breast Cancer is detected, it will provide the information/results in the following format and will store them in the Cloud database:

Breast Cancer Detected: Yes - Stage 0/1/2/3/4

If Breast Cancer is not detected, our application will invoke the Random Forest-based predictor that will help analyze the healthy patient's present-day medical parameters to determine their chances of developing Breast Cancer in the coming years. It will provide the percent chance of developing Breast Cancer in the following format and will store it in the Cloud database:

Percent Chance of Developing Breast Cancer: #%

(Here # is the number)

All medical parameters, reports and results associated with the patient will automatically be stored in the MongoDB Atlas (AWS Instance) based Cloud storage.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

C. DenseNet Algorithm for Breast Cancer Detection

1) Description

For our Detection module, we leverage the DenseNet201 classifier [25], a variant of the DenseNet (Dense Convolutional Neural Network) that consists of 201 layers, 4 DenseBlocks and 3 Transition Layers that help us detect the presence of Breast Cancer accurately. Here, we are using a Histopathological based Image dataset focusing on Invasive Ductal Carcinoma (IDC) [24]. It has Histopathological images associated with 735 such clinical cases.

We use exploratory data analysis to capture a better insight into the dataset by making a set of subplots that represent both cancerous and non-cancerous tissues of such patients. We further shuffle our Image dataset to enforce diversification, create random and unique samples. With our dataset ready, we then employ the technique of Cross-Validation to segment the dataset into 80% Train and 20% Test.

We build our DenseNet201 Convolutional Neural Network-based sequential model to help identify and detect the presence of Breast Cancer. DenseNet follows a feed-forward [26] architecture and ensures that the information flow remains maximum by connecting each layer with all previous layers. Here, each layer (L(i)) receives a collective knowledge of previous-layers and hence each layer provides its Feature Maps to the following L-1 layers, with direct connection consisting of (L(L+1))/2 layers.

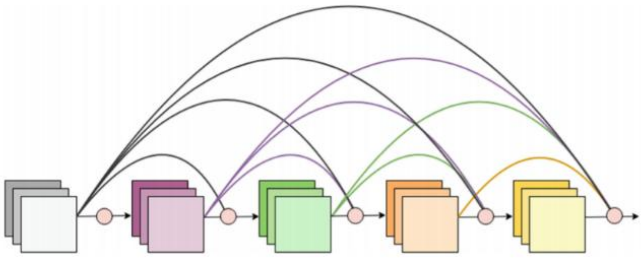


Figure 7. Collective Information Flow and Layer Connectivity [16] in DenseNet Network

DenseNet divides the entire network into multiple DenseBlocks, with each of them having a different set of "Fitters". Each DenseBlock further leverages a Composite Function [26] consisting of Batch Normalisation, ReLU and Convolution of 3x3.

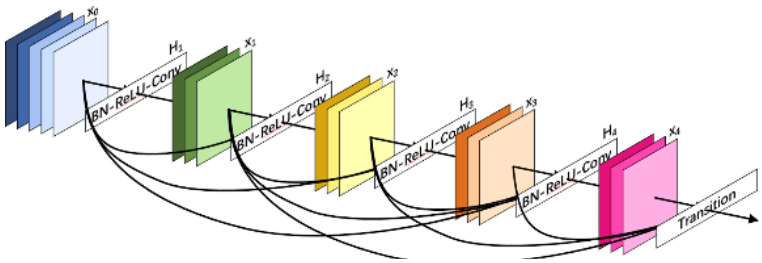


Figure 8. Composite Function [26] Between DenseNet Layers

We then use DenseNet to produce and transfer the collective knowledge of output Feature Maps by concatenating them using the following equation:

$$x_l = H_l([x_0, x_1, \ldots x_{l-1}]) \quad (3)$$

Equation (3) is the Mathematical Formula for Feature Map Information Concatenation.
Here,

- $H_{(l)}$ =It represents the Composite Function.
- $[X_{(0)}, X_{(1)}..]$=It represents the Feature Maps in a concatenated form

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Using DenseNet's Transition Layers, consisting of Convolution and Pooling Layer, we can downsample the output Feature Maps into the same dimensions and feed them to the next DenseBlock [27]. It helps ensure a robust gradient flow and efficient transfer of collective knowledge.

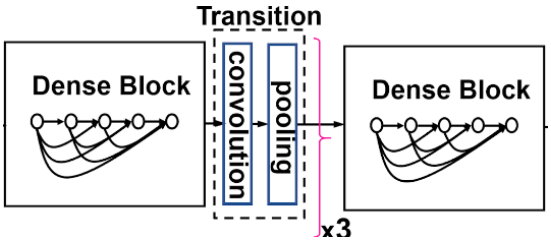


Figure 9. Visual Representation of Transition Layer [26] Structure in DenseNet

Each Convolution Layer that is a part of the Transition helps us to extracts the features from the previous-layer and provide them as the input to the next activation layer:

$$z^l = W^l . f_1\left(z^{(l-1)}\right) + b^l \qquad (4)$$

Equation (4) is the Mathematical Formula for Information Extraction by Convolution Layer.

Here,

- $\mathbf{W^l}$= It represents the weight matrix.
- $\mathbf{F_l}$= It represents the activation function.
- $\mathbf{Z^l}$= It represents the neutrons of the l^{th} layer.
- $\mathbf{B^l}$= It represents the bias associated with l^{th} to $(I-1)^{th}$ layers.

In our model, we use DenseNet as the backbone and set the HyperParameters such as Batch Size as 16, Learning Rate as 1e-4, Input Shape to (50,50,3), the Weights as ImageNet and the Loss as "cross-entropy". We further build the classification layer of our model by leveraging a GlobalAveragePooling, BatchNormalisation and Softmax Layer.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
densenet201 (Functional)     (None, 1, 1, 1920)        18321984
_________________________________________________________________
global_average_pooling2d (Gl (None, 1920)              0
_________________________________________________________________
dropout (Dropout)            (None, 1920)              0
_________________________________________________________________
batch_normalization (BatchNo (None, 1920)              7680
_________________________________________________________________
dense (Dense)                (None, 2)                 3842
=================================================================
Total params: 18,333,506
Trainable params: 18,100,610
Non-trainable params: 232,896
_________________________________________________________________
```

Figure 10. Description of DenseNet Model Summary

The GlobalAveragePooling Layer helps us unify and combine the spatial and statistical output from all the network layers and apply normalisation.

We then use the Softmax Classifier and provide it real-number based vector input that will help scale all values and bring the sum of the output values as 1, thereby considering them as probabilities associated with input params. Softmax Classifier uses the following formula:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} j = 1, \ldots., Kz = (z_1, \ldots, z_k) \in R^K \quad (5)$$

Equation(5) is the Mathematical Formula for Softmax Classifier.

Hence, In our project, by running our model for up to 20 epochs, we were able to achieve an accuracy of ~ 97.59%

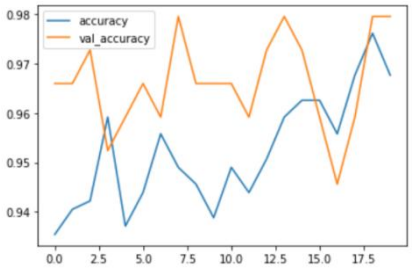


Figure 11. Accuracy Validation Curve for DenseNet

The Confusion Matrix of the same is as follows:

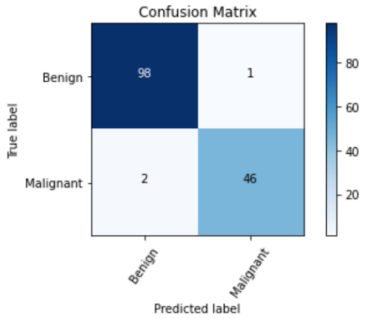


Figure 12. Confusion Matrix of DenseNet

We calculate the model accuracy by leveraging the True Positive and True Negative values from the Confusion Matrix as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{98+46}{98+1+46+2} = .9759 \quad (6)$$

Equation (6) is the Mathematical Formula for DenseNet Accuracy Calculation.

The Receiver Operating Curve (ROC) and the Area Under the Curve (AUC) associated with our model are as follows:

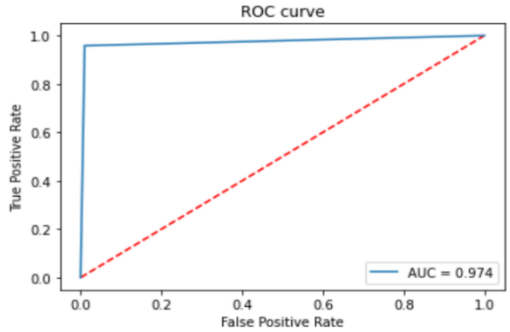


Figure 13. Receiver Operating Curve (ROC) for DenseNet

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

D. Random Forest Algorithm for Breast Cancer Prediction

1) Algorithm

- **Step 1:** Input the dataset and use the Cross-Validation technique to split the dataset into testing and training sets.
- **Step 2:** Randomly select N-data points from the desired dataset.
- **Step 3:** Leverage the N-data points (Subset) to create Decision Trees for each of them.
- **Step 4:** Repeat Step 2 & 3 and generate the Decision Tree for each new set of data points.
- **Step 5:** Determine the prediction of each set of data points.
- **Step 6:** Carry out of voting for each prediction result value.
- **Step 7:** Declare the value with the maximum number of votes as the final prediction value.

2) Description

For our Prediction module, we leverage the Random Forest Algorithm to provide an accurate prediction for chances of developing Breast Cancer. Here, we take the Wisconsin Breast Cancer (Diagnostic) dataset [23], consisting of values for 33 unique medical parameters for 569 clinical cases.

We carry out exploratory data analysis that helps understand how each feature is associated with the other. We do this by plotting a heatmap for these values. We then drop two columns and all the null values from the dataset to facilitate the training process.

With the dataset ready, we use Cross-Validation to partition the dataset as 70% Train and 30% Test. We then apply the Random Forest Algorithm to determine the probability or percent chance of developing Breast Cancer. The Random Forest Algorithm uses the technique of "Bagging & Feature Randomness" to select N-data points as subsets and creates numerous Decision Trees for all such subsets.

By using Bagging and Feature Randomness, it ensures that each Decision Tree takes up random and unique data points and features. It thus creates numerous uncorrelated and diverse trees to predict the outcome. Here, we calculate the feature importance by monitoring the value of node impurity decrease and the probability of reaching that node.
Here, **Node Probability** = Number of Samples that Reach the Node / Total Samples.

To do this for each node, we use the Gini Importance formula [21] as follows:

$$ni_j = w_{left(j)}C_{left(j)} - w_{right(j)}C_{right(j)} \qquad (7)$$

Equation (7) is the Mathematical Formula for Gini Importance Value.

Here,

- $\mathbf{ni_{(j)}}$= It represents the importance of a node j.
- $\mathbf{w_{(j)}}$=It represents the number of weighted samples that reach node j.
- $\mathbf{C_{(j)}}$ = It represents the impurity value of the node j.
- $\mathbf{left_{(j)}}$ =It is the child node from the left split of node j.
- $\mathbf{right_{(j)}}$ =It is the child node from the right split of node j.

Now, to determine the feature importance at the Decision Tree level [21], we use the following formula:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_j} \qquad (8)$$

Equation (8) is the Mathematical Formula for Feature Importance Value.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Here,

- **$fi_{(i)}$** = It represents the importance of each feature i.
- **$ni_{(i)}$** = It represents the importance of each node j.

To scale the values in a single range of 0-1, we divide the importance value of each feature by the sum of all "Importance" values as follows:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \quad (9)$$

Equation (9) is the Mathematical Formula for Normalization Function.

To poll the prediction values from all the Decision Trees and determine the final prediction value at the Random Forest level, we carry out a sum of all the feature importance values associated with each tree and then divide it by the count of all trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normf\ i_j}{T} \quad (10)$$

Equation (10) is the Mathematical Formula for Random Forest Prediction Value.

Hence, in our project, we were able to achieve a training time of 0.2 seconds and a prediction time of 0.015 seconds, along with an accuracy of ~ 97.07%.

```
Training time: 0.2 s
Prediction time: 0.015 s

Report:

Accuracy: 0.9707602339181286

              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98       108
         1.0       0.98      0.94      0.96        63

    accuracy                           0.97       171
   macro avg       0.97      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171

[[107   1]
 [  4  59]]
```

Figure 14. Random Forest Classification Report

The normalised Confusion Matrix of the same is as follows:

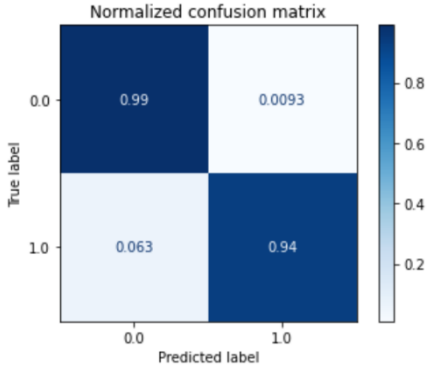


Figure 15. Confusion Matrix of Random Forest

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

We calculate the model accuracy by leveraging the True Positive and True Negative values from the Confusion Matrix as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{107+59}{107+1+59+4} = .9707 \qquad (11)$$

Equation (11) is the Mathematical Formula for Random Forest Accuracy Calculation.

The Receiver Operating Curve (ROC) and the Area Under the Curve (AUC) associated with our model are as follows:

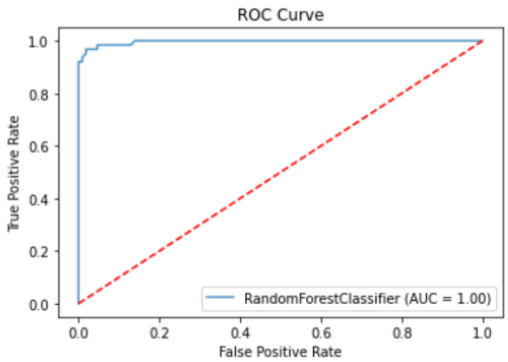


Figure 16. Receiver Operating Curve (ROC) for Random Forest

E. MongoDB Atlas Cloud Database (AWS Instance)

To power our "Integrated Breast Cancer Analyzer and Predictor" application with robust storage support, we use the MongoDB Atlas-based Cloud database as our primary data storage backend.We host our MongoDB NoSQL Cluster on the AWS Instance [28] that allows us to leverage it's fully-managed and robust processing power to store and fetch the patient records seamlessly. It further allows us to scale the storage requirements in real-time without hampering the performance of the application.

We have specifically chosen our database type to be a NoSQL database as it allows us to store the patient records without enforcing any table structure/schema constraints, thereby allowing us to store and process complex medical parameters associated with the patients.

It stores the data associated with the patient in the form of MongoDB Documents (Records), each having a unique ID that helps identify and fetch the patient details with ease. Each medical parameter and the test results associated with the patient are available and stored in the form of a key-value pair.

To connect our MongoDB Atlas-based NoSQL database cluster with our Python-based application, we use the PyMongo library [29] and establish a connection by creating a MongoClient for the same.

ResultAnalysis and Discussion

To measure & keep track of the performance of our robust "Integrated Breast Cancer Analyzer and Predictor" application, we leverage four essentials performance metrics. These parameters help provide holistic insights into the functioning, and accurate prediction & classification capabilities of our DenseNet201 powered Breast Cancer Detection module & Random Forest powered Breast Cancer Prediction module.

To calculate each performance indices, we leverage the Confusion Matrix [30], consisting of the True Positive, True Negative, False Positive & False Negative values.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

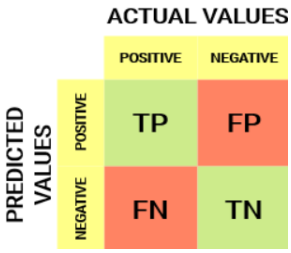


Figure17. Visualization of Generic Confusion [30] Matrix

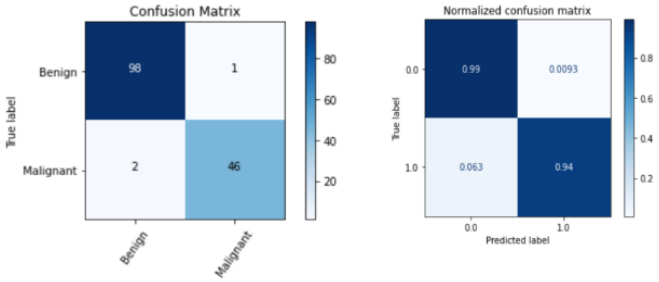


Figure18. Confusion Matrix for DenseNet and Random Forest

A. Accuracy

It is one of the most essential metrics that help gain detailed insights into the effectiveness & correctness of the training process of the module/s. Accuracy [31] helps understand the measure of extent up to which the module will be able to provide accurate and reliable results.

To calculate the accuracy of our Detection & Prediction modules, we leverage the True Positive, True Negative, False Positive & False Negative values taken up from the Confusion Matrices of each of them using the following formula:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

Equation (12) is the Mathematical Formula for Accuracy Calculation.

1) DenseNet201 Detection Module

Here, TP = 98, FP = 1, FN = 2, TN = 46. Using these values from our Confusion Matrix, we calculate the accuracy as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{98+46}{98+1+46+2} = .9759 \quad (13)$$

Equation (13) is the Mathematical Formula for DenseNet Accuracy Calculation.

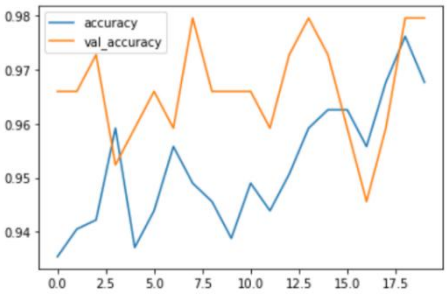


Figure 19. Accuracy Validation Curve for DenseNet

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Hence, we were able to achieve exceptionally high accuracy of ~ 97.59% for our DenseNet201 based Detection module, which can help reliably determine the presence of Breast Cancer by analysing the Histopathological images.

2) Random Forest Prediction Module

Here, TP = 107, FP = 1, FN = 4, TN = 59. Using these values from our Confusion Matrix, we calculate the accuracy as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{107+59}{107+1+59+4} = .9707 \qquad (14)$$

Equation (14) is the Mathematical Formula for Random Forest Accuracy Calculation.

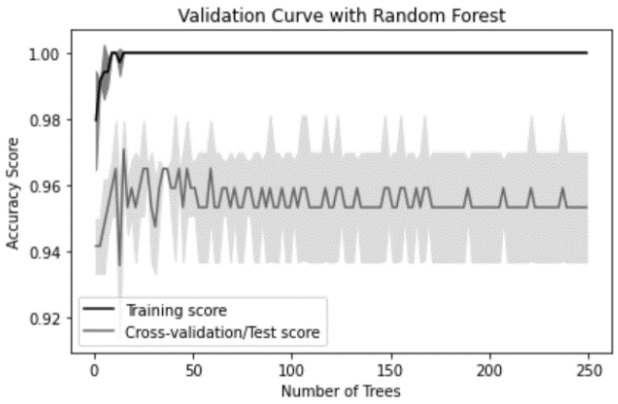


Figure 20. Accuracy Validation Curve for Random Forest

Hence, we were able to build an effective and reliable model, having an impressive accuracy of ~ 97.07%. Our highly accurate model can help gain in-depth insights about a presently healthy patient's chances of developing Breast Cancer in the coming years.

B. Precision

It plays a vital role in understanding the degree of correctness of the predicted outcomes of a module. Precision [31] values represent the accurate count of predicted cases that turned out to be true. It also helps provide holistic insights into the handling capacity of the module in terms of the positive cases.

To determine the precision of our Detection & Prediction modules, we leverage the True Positive and False Positive values taken up from the Confusion Matrices using the following formula:

$$Precision = \frac{TP}{TP+FP} \qquad (15)$$

Equation (15) is the Mathematical Formula for Precision Calculation.

1) DenseNet201 Detection Module

Here, TP = 98 and FP = 1. We leverage these values to calculate the precision for our DenseNet201 powered Detection module as follows:

$$Precision = \frac{TP}{TP+FP} = \frac{98}{98+1} = .9787 \qquad (16)$$

Equation (16) is the Mathematical Formula for DenseNet Precision Calculation.

Hence, 97.87% of the cases detected by our robust Detection module turned out to be positive.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

2) Random Forest Prediction Module

Here, TP = 107 and FP = 1. We leverage these standard values from the Prediction module's Confusion Matrix to calculate the precision as follows:

$$Precision = \frac{TP}{TP+FP} = \frac{107}{107+1} = .9833 \quad (17)$$

Equation (17) is the Mathematical Formula for Random Forest Precision Calculation.

Hence, we were able to achieve an excellent precision value of 98.33% for our Random Forest powered Prediction module.

C. Recall

It is one of the most essential metrics that help understand the ability and efficiency of the module in terms of detecting the positive values. It can be represented as the ratio of the observations/classifications that are correctly determined to the total number of observations. It acts as the measure of the sensitivity of any system; and helps provide a holistic insight into the count of real positive cases that a module was able to detect correctly, thereby making one of the essentials performance indices.

To calculate the Recall [31] value, we leverage the True Positive and False Negative values taken up from the Confusion Matrices using the following formula:

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

Equation (18) is the Mathematical Formula for Precision Calculation.

1. DenseNet201 Detection Module

Here, TP = 98 and FN = 2. By using these values from the associated Confusion Matrix, we determine the Recall value for our DenseNet201 based Detection module as follows:

$$Recall = \frac{TP}{TP+FN} = \frac{98}{98+2} = .9583 \quad (19)$$

Equation(19) is the Mathematical Formula for DenseNet Recall Calculation.

Hence, we were able to obtain a Recall value of ~ 95.83%, indicating that our model was able to detect nearly 95% of truly positive values successfully.

2. Random Forest Prediction Module

Here, TP = 107 and FN = 4. We leverage these values to determine the Recall value as:

$$Recall = \frac{TP}{TP+FN} = \frac{107}{107+4} = .9356 \quad (20)$$

Equation (20) is the Mathematical Formula for Random Forest Recall Calculation.

Hence, we were able to obtain a Recall value of ~ 93.65%, indicating that our model was able to predict nearly 94% of truly positive values successfully.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

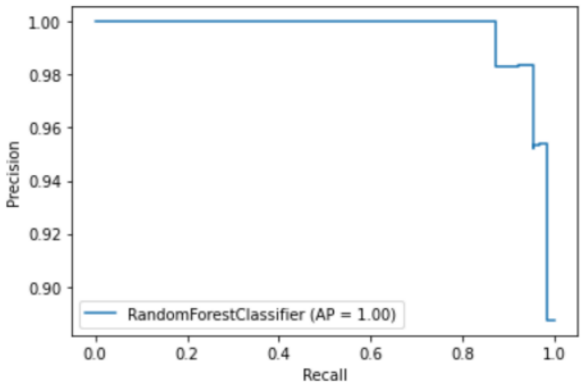


Figure 21. Precision vs Recall (PR) Curve for Random Forest

D. F-Score

The F-Score [31] value helps provide an integrated insight into Precision and Recall trends. Values for the F-Score ranges between 0-1, with 1 being the perfect value and 0 being the failure value. It can be calculated as the harmonic mean of both Recall & Precision values for your module/s. A perfect value of F-Score (1) is achieved when you have Recall value = Precision value for your Machine Learning based module.

To calculate the F-Score for our Detection and Prediction Modules, we leverage the associated Recall and Precision values as follows:

$$F - Score = \frac{2*Precision\ *Recall}{Precision\ +Recall} \qquad (21)$$

Equation (21) is the Mathematical Formula for Precision Calculation.

1) DenseNet201 Detection Module

Here, Recall = 95.83 and Precision = 97.87. By leveraging the Recall & Precision values for the DenseNet201 powered Detection module, we calculate the F-Score as follows:

$$F - Score = \frac{2*Precision\ *Recall}{Precision\ +Recall} = \frac{2*97.87*95.83}{97.87+95.83} = .968 \qquad (22)$$

Equation (22) is the Mathematical Formula for DenseNet F-Score Calculation.

Hence, we were able to achieve an F-Score of ~ 96.82% for our Detection module.

2) Random Forest Prediction Module

Here, Recall = 93.65 and Precision = 98.33. We use the Precision and Recall values associated with our Random Forest algorithm-based prediction values as follows:

$$F - Score = \frac{2*Precision\ *Recall}{Precision\ +Recall} = \frac{2*98.33*93.65}{98.33+93.65} = .959 \qquad (23)$$

Equation (23) is the Mathematical Formula for Random Forest F-Score Calculation.

Hence, we were able to achieve an F-Score of ~ 95.93% for our Prediction module.

The summarised results for our DenseNet powered Detection module & Random Forest Powered Prediction module are as follows:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

Table 3. Summarised Values for Performance Metrics

Metric/Algorithm	DenseNet	Random Forest
Accuracy	97.95%	97.07%
Precision	97.87%	98.33%
Recall	95.83%	93.65%
F-Score	96.84%	95.93%

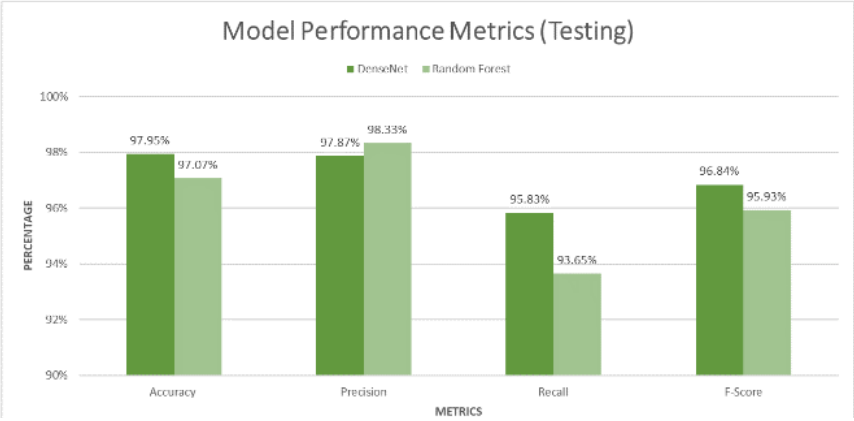


Figure 22. Graphical Representation of Performance Metrics

Conclusion

In our "Integrated Breast Cancer Analyzer and Predictor" application, we leveraged two main algorithms that are DenseNet and Random Forest. We applied the DenseNet201 Convolutional Neural Network on the Histopathological Images-based Invasive Ductal Carcinoma (IDC) dataset and found that we were able to build a reliable and highly accurate model that could detect the presence of Breast Cancer with high accuracy of 97.59%

We further integrated and invoked our Prediction module for patients who would test healthy as per our DenseNet201-based module. We leveraged the Random Forest classifier to power our Prediction module and applied it to the Wisconsin Breast Cancer (Diagnostic) dataset and found that we were able to build a robust and accurate model that could analyze the healthy patient's present medical parameters and predict their chances of developing Breast Cancer with high accuracy of 97.07%

The results thus obtained indicate that the integrated application of Supervised Machine Learning algorithms such as Random Forest and Deep Learning approaches such as DenseNet can be of great support in early-stage prognosis and diagnosis of Breast Cancer among women and cancer research as well. Further, these can not only be a great boon for accurately detecting & predicting Breast Cancer but also help spread awareness among women, thereby allowing them to get timely treatment and potentially save their lives. This approach further proves that an effective diagnosis can be made without having exceptional medical bandwidth & expertise.

Future work can be carried out to enhance these integrated approaches by applying them across more complex and large datasets and carrying out deep Multi-Class classification beyond the Cancer stage level.

References

[1] “About Breast Cancer”, American Cancer Society, Cancer.org. (https://www.cancer.org/content/dam/CRC/PDF/Public/8577.00.pdf).

[2] "Breast Anatomy", Cleveland Clinic. (https://my.clevelandclinic.org/health/articles/8330-breast-anatomy).

[3] R.Saalovra., L.A. Altonen, P. Kristo, F. Canzian, A. Hemminki, Peltomaki P, R. Chadwik, A. De La Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease", *N Engl J Med.*, Vol. 337, pp. 1481–1487, 1998.

[4] J. Weston, I. Guyon, S. Barnhill, V. Vapnik. "Gene selection for cancer classification using support vector machines". *Machine Learning*, Vol. 46, pp. 389–422, 2002.

[5] Nitesh Pradhan, Tanishk Thomas, Vijaypal Singh Dhaka. "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey". February 2020, DOI: 10.1109/ICICT48043.2020.9112464.

[6] Quang H. Nguyen, Trang T.T. Do, Yijing Wang, Sin Swee Heng, Kelly Chen, Wei Hao Max Ang, Conceicao Edwin Philip, Misha Singh. *"Breast Cancer Prediction using Feature Selection and Ensemble Voting"*. July 2019, DOI: 10.1109/ICSSE.2019.8823106.

[7] H. Zhang, Y. Jiang, L. C. Id and X. Xiao, "Breast Cancer histopathological image classification using convolutional neural networks with small SE-ResNet module," pp. 1–21, 2019.

[8] Dalal Bardou, Kun Zhang, Sayed Mohammad Ahmad. "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Network". May 2018, DOI: 10.1109/ACCESS.2018.2831280.

[9] S. Pal, V. Chaurasia, and B. B. Tiwari, “Prediction of Benign and Malignant Breast Cancer using Data Mining Techniques,” 2018.

[10] R. Sangeeth, K. Srikanta Murthy. "A Novel Approach for Detection of Breast Cancer at an Early Stage using Digital Image Processing Techniques". January 2017, DOI: 10.1109/ICISC.2017.8068625.

[11] K. S. Sim, Y. J. Tan, and F. F. Ting, “Breast Cancer Detection Using Convolutional Neural Networks for Mammogram Imaging System.” 2017.

[12] Ladislav Lenc and Pavel Král. "LBP features for breast cancer detection". September 2016, DOI: 10.1109/ICIP.2016.7532838.

[13] Mahersia Hela, Boulehmi Hela, Hamrouni Kamel, Boussetta Sana, Mnif Najla. "Breast Cancer Detection: A Review on Mammograms Analysis Techniques". March 2013, DOI: 10.1109/SSD.2013.6563999.

[14] Daniel R.Bauer, Xiong Wang, Russell Witte and Hao Xin. "Microwave-Induced Thermoacoustic Imaging Model for Potential Breast Cancer Detection". July 2012, DOI: 10.1109/TBME.2012.2210218.

[15] "A Comprehensive Guide to Convolutional Neural Networks". Towards Data Science.

(https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53).

[16] Anita Rybiałek and Lukasz Jele. "Application of DenseNet for Classification of Breast Cancer Mammograms". May 2020, DOI: 10.1007/978-3-030-47679-3_23.

[17] "DenseNet — Dense Convolutional Network (Image Classification)". Towards Data Science.

(https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803).

[18] "Understanding and visualizing DenseNets". Towards Data Science.

(https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a)/

[19] "A Complete Guide to Random Forest Algorithms". Built.in

(https://builtin.com/data-science/random-forest-algorithm).

[20] "Understanding Random Forest". Towards Data Science.

(https://towardsdatascience.com/understanding-random-forest-58381e0602d2).

[21] “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark”. Towards Data Science.

(https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4460 - 4478
Received 25 April 2021; Accepted 08 May 2021.

[22] Amrane Meriem and Ikram Gagaoua. "Breast Cancer Classification using Machine Learning". April 2018, DOI: 10.1109/EBBT.2018.8391453.

[23] Z. Zhou, Y.J., Y. Yang, S.F. Chen, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence", *Medicine Elsevier,* Vol. 24, pp. 25-36, 2002.

[24] Barath Narayanan Narayanan, Vignesh Krishnaraja and Redha Ali. "Convolutional Neural Network Classification of Histopathology Images for Breast Cancer Detection". July 2019, DOI: 10.1109/NAECON46414.2019.9058279.

[25] Feyza Yılmaz, Onur Konse and Ahmet Demir. "Comparison of Two Different Deep Learning Architectures on Breast Cancer". October 2019, DOI: 10.1109/TIPTEKNO47231.2019.8972042

[26] Xai Shen, Xia Li, Xiuhui Wang and Yongxia Zhou. "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)". May 2020, DOI: 10.1371/journal.pone.0232127.

[27] "Introduction to DenseNet with TensorFlow". Pluralsight.

(https://www.pluralsight.com/guides/introduction-to-densenet-with-tensorflow).

[28] “MongoDB Atlas with AWS”. MongoDB Atlas.

(https://www.mongodb.com/cloud/atlas/aws-mongodb).

[29] “pymongo Library”. pymongo. https://pypi.org/project/pymongo/.

[30] "Everything you Should Know about Confusion Matrices for Machine Learning". Analytics Vidhya.

(https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/).

[31] Shubham Sharma, Archit Aggarwal and Tanupriya Choudhury. "Breast Cancer Detection Using Machine Learning Algorithms". December 2018, DOI**:** 10.1109/CTEMS.2018.8769187.0232127.