iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35107

www.ijraset.com

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

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

Breast Cancer Prediction using CNN and Machine Learning Algorithms with Comparative Analysis

Santhosh Voruganti[1], U. Sairam[2]
[1, 2]*Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, India-500075*

Abstract: ***At present world, Breast cancer is a second main cause of cancer death in women after lung cancer. Breast cancer occurs when some breast cells begin to raise abnormally. It can arise in any portion of the Breast and it can be prevented if the treatment is started at the early stage of the Breast cancer. Breast cancer is a malignant tumour i.e. a collection of cancer cells arising from the cells of the breast Treatment of breast cancer relies on the cancer type and its stage. Mainly this paper focused on diagnosing the Breast cancer disease using various classification algorithm with the help of data mining tools. Data mining of the intelligent accumulated from previously disease detected patients opened up a new aspect of medical progression In this paper, the focus has been prediction of breast cancer using various machine learning algorithms and visualizing the performances of each algorithm. This paper makes use of a dataset that contains numerical values about the clump thickness, uniformity of the cell for prediction using Multi Layer Perceptron, K-NN, Random Forest, Logistic Regression. Moreover, this also uses a dataset of tissue images for the prediction using Convolution Neural Network. Later, the accuracies of each algorithm is calculated along with the precision, recall, f-score, ROC for each algorithm.***
Keywords: Multi Layer Perceptron, K-NN, Naive Bayes Classifier, Logistic Regression, Convolution Neural Network

I. INTRODUCTION

A. Overview

Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumour. A tumour can be cancerous or benign. A cancerous tumour is malignant, meaning it can grow and spread to other parts of the body. A benign tumour means the tumour can grow but will not spread Breast cancer spreads when the cancer grows into other parts of the body or when breast cancer cells move to other parts of the body through the blood vessels and/or lymph vessels. It can arise in any portion of the Breast and it can be prevented if the treatment is started at the early stage of the Breast cancer. Breast cancer is a malignant tumour is a collection of cancer cells arising from the cells of the breast Treatment of breast cancer relies on the cancer type and its stage (zero to fourth) and may include surgery, radiation, or chemotherapy. The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci, the prediction of cancer susceptibility (risk assessment) the prediction of cancer recurrence and the prediction of cancer survivability. Breast Cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body.

B. Motivation

Now a day, breast cancer is one of the burning issue all over the world. It is one of the major health problem for women. Globally the incidence of breast cancer is only second to that of Lung cancer. The disease represents the main cause of cancer death among women. Breast cancer is developed from breast tissue. Signs of breast cancer may include a breast lump, skin dimpling, fluid coming from the nipple, breast shape change, a newly inverted nipple, or a scaly patch of skin. Breast cancer typically attack postmenopausal women. Both genetic and ancestral factor play a role. Prolong estrogen exposure associated with early menarche, late menopause uses of hormone replacement therapy has been associated with increased risk. Other risk factor includes Obesity Alcohol intake, nulliparity and late first pregnancy. Breast cancer usually present as a Palpable mass with nipple discharge. Breast cancer may metastasis to bone, lung, liver and other organs. Breast Cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body.

C. Applications

There are many number of application for breast cancer prediction. Few of the applications are explained below
The prediction of breast cancer is used to know whether there is cancer or not and its stage (benign or malignant).
The patient who is having breast cancer can undergo treatment where they can be get cured.
The approximate the life span of a person can also be depicted based on the stage.
It is mainly the use of machine learning algorithms for breast cancer prediction, thus providing accurate information about the person's status whether having cancer or not.

D. Problem Statement

In the real world, prediction of breast cancer manually by doctors can cause some errors to occur unintentionally where the results may not be accurate or correct. Wrong prediction about the cancer is very risky where the life of a person can be in danger. In order to solve this problem machine learning algorithms are used to predict the breast cancer. The accuracy that is calculated can while predicting tells its performance.

E. Objective

The main objective of our project is prediction of breast cancer and also showing its stage. The paper deals with many machine learning algorithms such as Multi Layer Perceptron, K-NN, Random Forest , Logistic Regression and also Convolution Neural Network. In this paper, accuracy values are also calculated to visualize the performance of each algorithm.

II. RELATED WORK

A deep neural network (DNN) is used for predicting the prognosis of human breast cancer. The DNN architectures build a hierarchy from the hidden layers. Higher level features are extracted implicitly by the combination of lower-level features from each layer. Here, a DNN model is composed of an input layer, multiple hidden layers and an output layer. Units between layers are all fully connected. The input layer with an input vector x consists of one or multi-dimensional data. The output for layer k including j units is calculated from the weighted sum of the outputs for the previous layer. Afterwards, we initialize the weights between each layer using normalized initialization. One of the most straightforward approaches for discriminative tasks is to train only one DNN model for all multi-dimensional data. However, different data may have different feature representation, and directly combining the three sources of data as an input of a DNN model may not be efficient. We address this problem by proposing a multimodal DNN model which efficiently integrates multi-dimensional data.
The proposed system have a novel Multimodal Deep Neural Network by integrating Multi-dimensional Data (MDNNMD) for human breast cancer prognosis prediction. MDNNMD is an efficient method to integrate multi-dimensional data including gene expression profile, copy number alteration (CNA) profile and clinical data with a score level fusion at the final prediction results. This method considers the heterogeneity among different data types and makes full use of abstract high-level representation from each data source. The results of ten-fold cross validation experiment show that MDNNMD achieves an overall better performance than the prediction methods with single-dimensional data and existing research approaches: support vector machine (SVM), random forest (RF) and logistic regression (LR). We also demonstrate the feasibility of the multimodal deep neural network and the usefulness of the multi-dimensional data in breast cancer prognosis prediction.

III. IMPLEMENTATION

Classification can be defined as the grouping of things by shared features, characteristics and qualities or if you will simply be dropping things into corresponding buckets, you could for instance classify the following geometric shapes based on their similarity. Classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.
Neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

We have the types of classification algorithms in Machine Learning

1) Logistic Regression
2) Naive Bayes Classifier
3) K-Nearest Neighbor
4) Neural Networks

A. Convolutional

Convolutional neural network layer types mainly include three types, namely Convolutional layer, pooling layer and fully-connected layer Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli. Each convolutional neuron processes data only for its receptive field. Although fully connected feed forward neural networks can be used to learn features as well as classify data, it is not practical to apply this architecture to images. A very high number of neurons would be necessary, even in a shallow (opposite of deep) architecture, due to the very large input sizes associated with images, where each pixel is a relevant variable. For instance, a fully connected layer for a (small) image of size 100 x 100 has 10000 weights for each neuron in the second layer. The convolution operation brings a solution to this problem as it reduces the number of free parameters, allowing the network to be deeper with fewer parameters. For instance, regardless of image size, tiling regions of size 5 x 5, each with the same shared weights, requires only 25 learnable parameters. In this way, it resolves the vanishing or exploding gradients problem in training traditional multi-layer neural networks with many layers by using back propagation. The aim of Convolutional layer is to learn feature representations of the inputs. Convolutional layer is consisting of several feature maps. Each neuron of the same feature map is used to extract local characteristics of different positions in the former layer, but for single neurons, its extraction is local characteristics of same positions in former different feature map. In order to obtain a new feature, the input feature maps are first convolved with a learned kernel and then the results are passed into a nonlinear activation function.

B. Pooling

Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at one layer into a single neuron in the next layer. For example, max pooling uses the maximum value from each of a cluster of neurons at the prior layer Another example is average pooling, which uses the average value from each of a cluster of neurons at the prior layer. The sampling process is equivalent to fuzzy filtering. The pooling layer has the effect of the secondary feature extraction, it can reduce the dimensions of the feature maps and increase the robustness of feature extraction. It is usually placed between two Convolutional layers. The size of feature maps in pooling layer is determined according to the moving step of kernels. The typical pooling operations are average pooling and max pooling. We can extract the high-level characteristics of inputs by stacking several Convolutional layer and pooling layer.

C. Fully connected

Fully connected layers connect every neuron in one layer to every neuron in another layer. It is in principle the same as the traditional multi-layer perceptron neural network (MLP).

The flattened matrix goes through a fully connected layer to classify the images. In general, the classifier of Convolutional neural network is one or more fully-connected layers. They take all neurons in the previous layer and connect them to every single neuron of current layer. There is no spatial information preserved in fully-connected layers. The last fully-connected layer is followed by an output layer. For classification tasks, SoftMax regression is commonly used because of it generating a well-performed probability distribution of the outputs. Another commonly used method is SVM, which can be combined with CNNs to solve different classification tasks.

D. Receptive field

In neural networks, each neuron receives input from some number of locations in the previous layer. In a fully connected layer, each neuron receives input from every element of the previous layer. In a convolutional layer, neurons receive input from only a restricted subarea of the previous layer. Typically the subarea is of a square shape (e.g., size 5 by 5). The input area of a neuron is called its receptive field. So, in a fully connected layer, the receptive field is the entire previous layer. In a convolutional layer, the receptive area is smaller than the entire previous layer.

E. Weights

Each neuron in a neural network computes an output value by applying some function to the input values coming from the receptive field in the previous layer. The function that is applied to the input values is specified by a vector of weights and a bias (typically real numbers). Learning in a neural network progresses by making incremental adjustments to the biases and weights. The vector of weights and the bias are called a filter and represents some feature of the input. A distinguishing feature of CNNs is that many neurons share the same filter. This reduces memory footprint because a single bias and a single vector of weights is used across all receptive fields sharing that filter, rather than each receptive field having its own bias and vector of weights.

F. Dataset

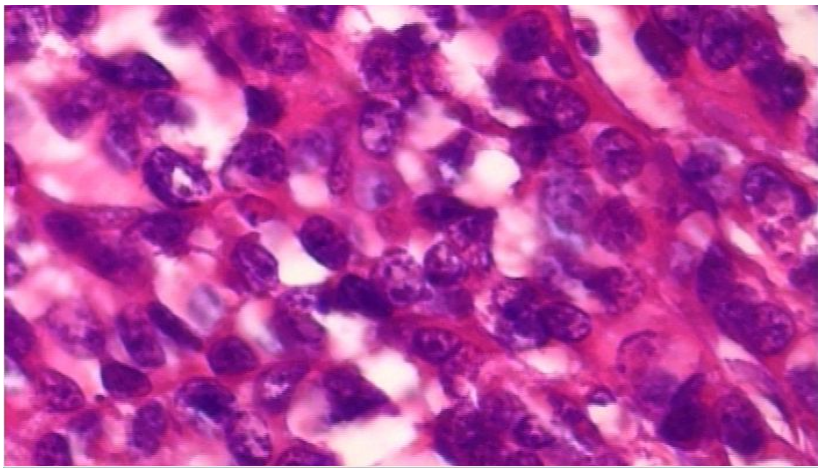


Fig.1 Tissue Image used in CNN

```
#Initialize the CNN
classifier = Sequential()
#Convolution and Max pooling
classifier.add(Conv2D(32, (2, 2), input_shape = (128,128, 3),activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2,2)))
classifier.add(Conv2D(64, (2, 2), activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2,2)))
classifier.add(Conv2D(128, (2, 2),activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2,2)))

#Flatten
classifier.add(Flatten())

#Full connection
classifier.add(Dense(128, activation = 'relu'))
classifier.add(Dense(2, activation = 'softmax'))

#Compile classifier
classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
```

Fig.2. Building the Model

```
#Fitting CNN to the images
train_datagen = ImageDataGenerator(rescale=1./255, shear_range=0.2, zoom_range=0.2, horizontal_flip=True)
test_datagen = ImageDataGenerator(rescale=1./255)
training_set = train_datagen.flow_from_directory('./Dataset/training', target_size=(128,128), batch_size=32, class_mode='categorical')
test_set = test_datagen.flow_from_directory('./Dataset/testing', target_size=(128,128), batch_size=32, class_mode='categorical')
classifier.fit_generator(training_set, steps_per_epoch=800/32, epochs=50, validation_data=test_set, validation_steps = 200/32)

#save model
classifier.save('model_cnn.h5')
classifier.save_weights('weights_cnn.h5')
```

Fig.3. Training the Model

```
#Prediction on a new picture
from keras.preprocessing import image as image_utils

from PIL import Image, ImageTk

class_labels = ['Benign', 'Malignant']
test_image = image.load_img('test2.png', target_size = (128, 128))
test_image = image.img_to_array(test_image)
test_image = np.expand_dims(test_image, axis = 0)

test_image /= 255
result = model.predict(test_image)

decoded_predictions = dict(zip(class_labels, result[0]))

decoded_predictions = sorted(decoded_predictions.items(), key=operator.itemgetter(1), reverse=True)
print("Predicted = ",decoded_predictions[0][0])


count = 1
for key, value in decoded_predictions[:5]:
    print("{}. {}: {:8f}%".format(count, key, value*100))
    count+=1
```

Fig.4. CNN prediction

Model. Summary() is used to see all parameters and shapes in each layers in our models. The total parameters are 3,728,354 and total trainable parameters are 3,728,354 with 0 Non-trainable parameters.

```
In [5]: classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])

In [6]: classifier.summary()
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 127, 127, 32)      416
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 63, 63, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 62, 62, 64)        8256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 31, 31, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 30, 30, 128)       32896
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 15, 15, 128)       0
_________________________________________________________________
flatten (Flatten)            (None, 28800)             0
_________________________________________________________________
dense (Dense)                (None, 128)               3686528
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 258
=================================================================
Total params: 3,728,354
Trainable params: 3,728,354
Non-trainable params: 0
_________________________________________________________________

In [ ]:
```

Fig.5. Model Summary

IV.RESULTS

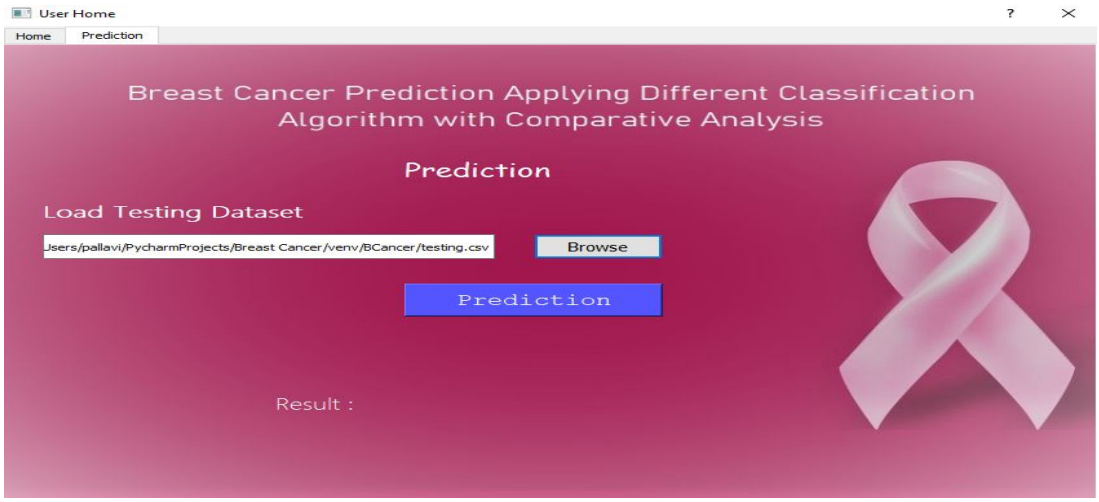


Fig.6 Loading the Test Data

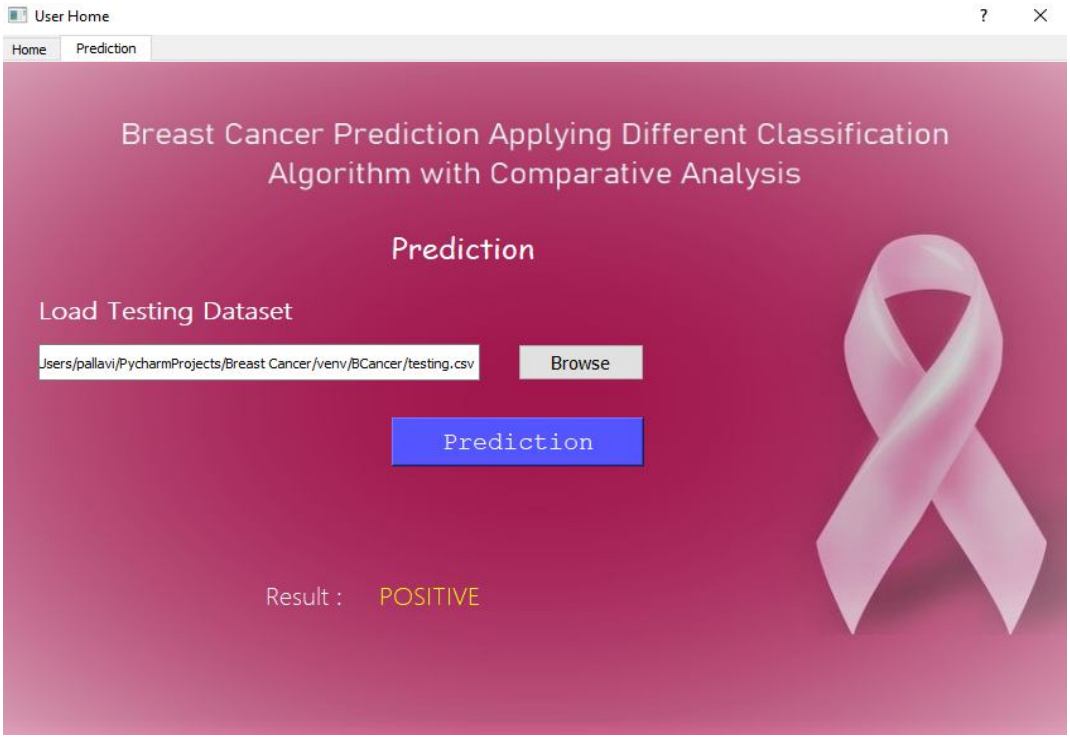


Fig.7. Prediction of Results

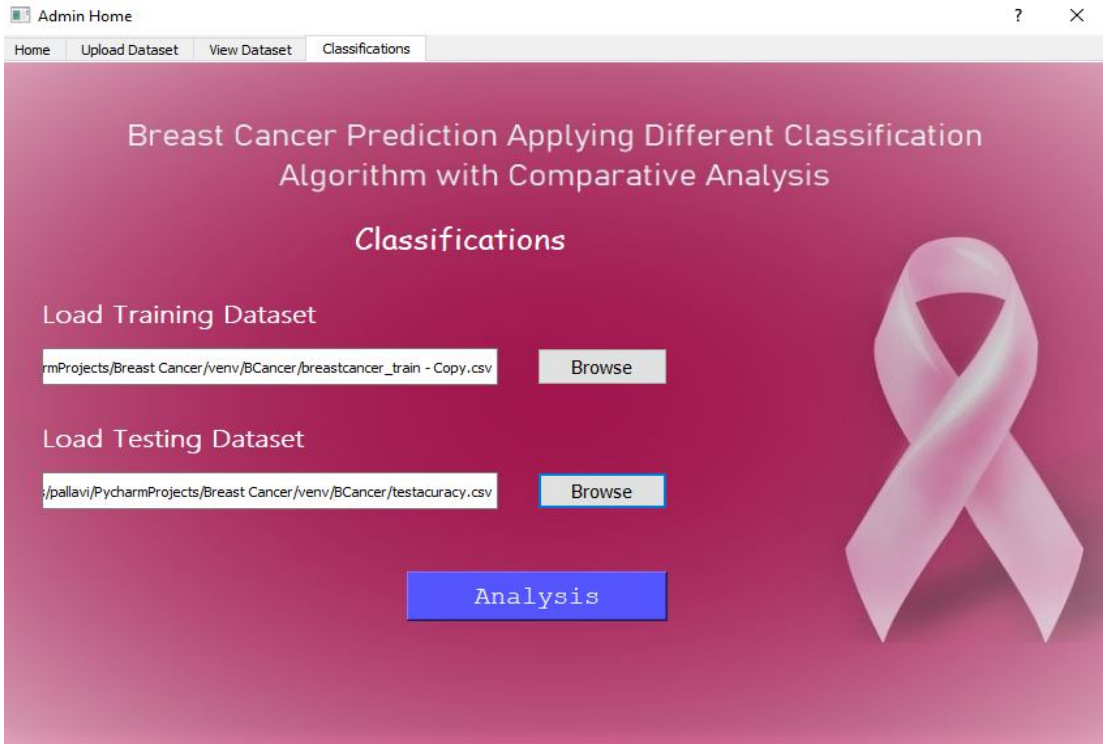


Fig.8 Loading dataset for analysis

In the above figure ,we upload the testing and training datasets in order to analyze the working of algorithms. The algorithms that are included here are logistic regression K-Nearest Neighbor, Naive Bayes and Neural Networks. The accuracy for each algorithm is calculated and a graph is presented in order to visualize the performance of each algorithm.

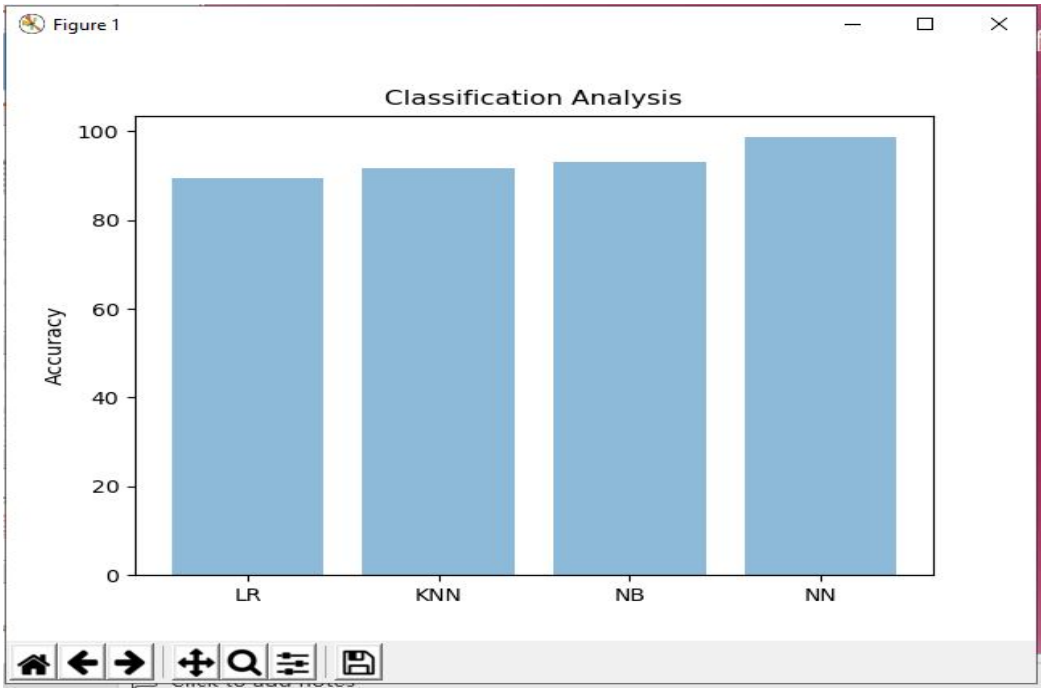


Fig.9. Graph of Accuracies for each algorithm

alg	accuracy	precision	recall	fScore	roc
KNN	91.66666666666666	0.9432624113475178	0.8866666666666667	0.9140893470790379	0.9166666666666667
LR	89.33333333333333	0.9682539682539683	0.8133333333333334	0.8840579710144927	0.8933333333333334
NB	93.0	0.9777777777777777	0.88	0.9263157894736842	0.9299999999999999
NN	98.66666666666667	1.0	0.9733333333333334	0.9864864864864865	0.9866666666666667

Fig.10 Tabular Comparison

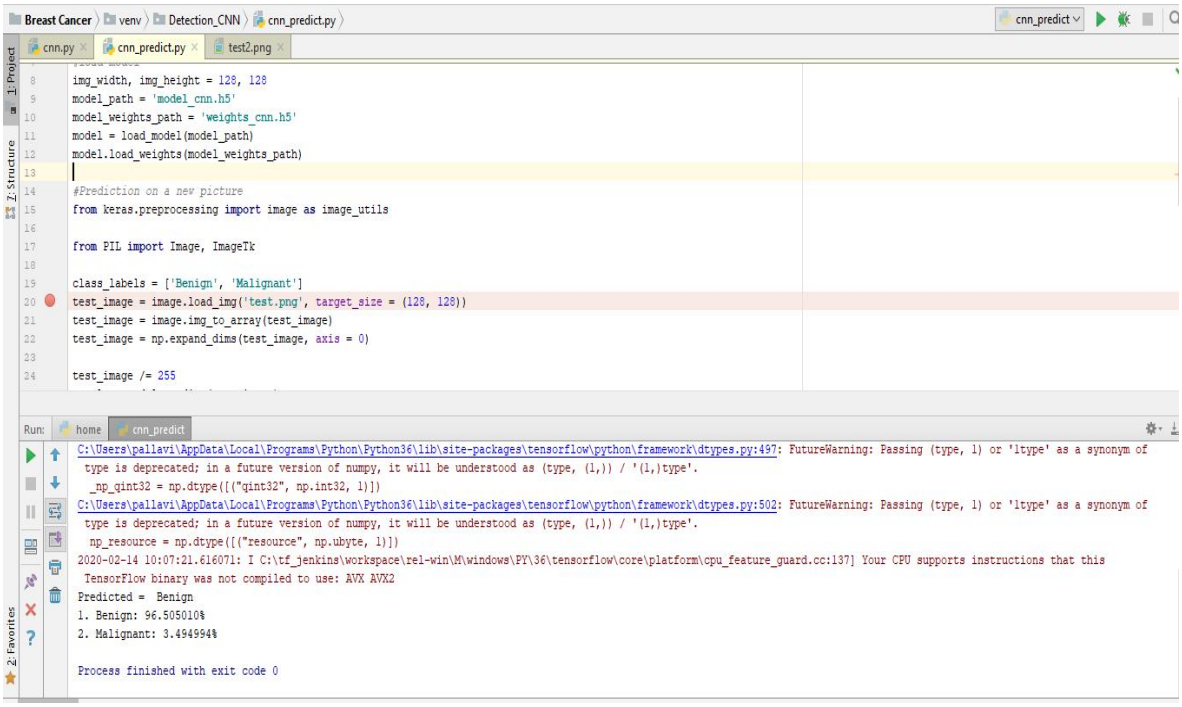


Fig.11 Prediction using CNN

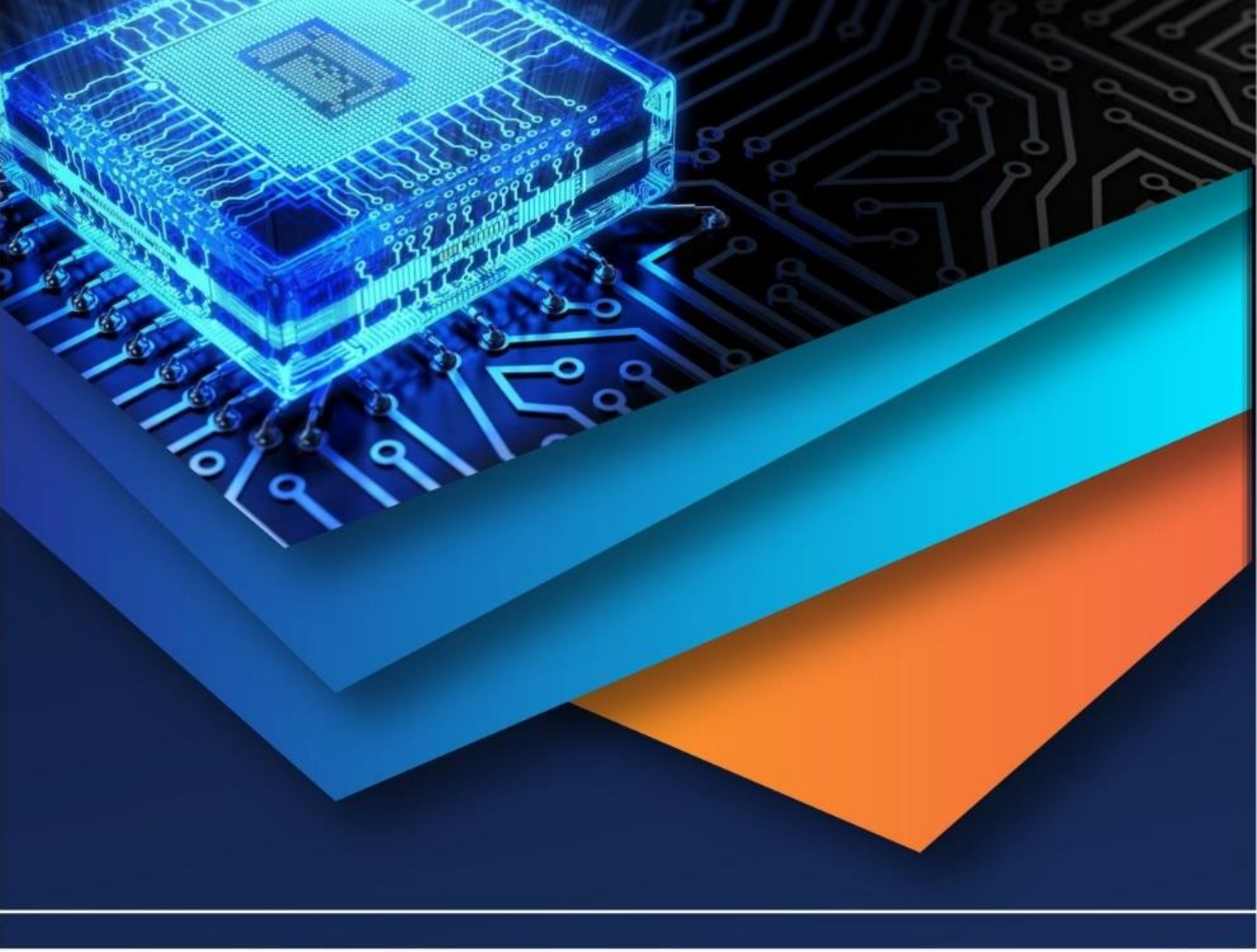
V. CONCLUSION AND FUTURE SCOPE

Breast Cancer has been predicted using the popular Data mining algorithms Naive Bayes, Logistic Regression, Multilayer Perceptron and K-nearest neighbours classifier using a popular dataset of 9 essential attributes. The results have been studied and used to compare the efficacy of the classifier algorithms based on Accuracy, Precision, Recall, F-Score and ROC. In terms of Accuracy, MLP or the Multilayer Perceptron was observed to perform the best with Logistic Regression having the least accuracy. Upon, analysis of all classification metric values between the algorithms, MLP outperforms all the other algorithms. It is seen that CNN, though gives high accuracy is actually difficult to interpret because the details of the operation (the algorithm) is not always understandable. In other words, if somebody asks as to how it is actually malignant one can't really answer. It is also seen that the data used is relatively small, about 446 collected images of tissue cells from patients for CNN and the database of about 1000 entries for the other four algorithms. Algorithms developed from such smaller databases would have limited validation and precision. Since our database has predominantly large number of mammograms with cancer, there will be a potential data bias in outcome while detecting cancer for the general population since breast cancer patients are much lesser.

Future scope of our paper is that we would love to implement tools which can be used to reduce the noise in images and improve the clarity and contrast. Currently medical image processing is one of the fastest growing areas in health care sector. As more GPU memory becomes available the performance can be increased as higher resolution is employed.

REFERENCES

[1] J. Ferlay, C. Héry, P. Autier, and R. Sankaranarayanan, "Global burden of breast cancer,in Breast cancer epidemiology, ed: Springer, Vol 20 Pages,2010.
[2] M. T. Islam, B. M. N. Karim Siddique, S. Rahman and T. Jabid, "Image Recognition with Deep Learning," International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Bangkok, pp. 106-110,2018.
[3] D. Q. T. Le, S. N. Tiwari, and B. Merialdo, "Deep Learning Image Recognition", Vol 1, 2015.
[4] Shomona G. Jacob and R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012, Vol. I, October 2012.
[5] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1-4,2018.
[6] Santhosh Voruganti Map reduce A programming model for cloud computing based on hadoop ecosystem published in International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3794-3799.
[7] Santhosh Voruganti Survey on Data-intensive Applications, Tools and Techniques for Mining Unstructured Data. International Journal of Computer Applications (0975 – 8887), volume 146-No.12, July 2016.
[8] Santhosh Voruganti Comparative Analysis of Dimensionality Reduction Techniques for Machine Learning IJSRST Volume 4 Issue 8 Print ISSN: 2395-6011 Online ISSN: 2395-602X Themed Section: Science and Technology June 2018.
[9] Santhosh Voruganti Enhanced Rating Prediction Based On Location And Friend Set published in JETIR May 2019 volume 6 issue 5 ISSN-2349-5162.
[10] Santhosh Voruganti Local Security Enhancement and Intrusion Prevention in Android Devices published in International Research Journal of Engineering and Technology Volume: 07 Issue: 01 January 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072.
[11] Santhosh Voruganti EFFECTIVE IOT TECHNIQUES TO MONITOR THE LEVELS OF GARBAGE IN SMART DUSTBINS published in International Research Journal of Engineering and Technology Volume: 07 Issue: 06 June 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072.
[12] U. Sairam Vanitha Kunta, Haritha Tuniki ,Multi-Functional Blind Stick for Visually Impaired People, Fifth International Conference on Communication and Electronics Systems, July 2020.
[13] Yashwant Adepu, Vishwanath R Boga, U Sairam, Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition,2020 IEEE International Conference for Innovation in Technology (INOCON), pages 1-5 in 06.11.2020.
[14] M V Bhanu Prakash U Sairam, Feature Prospect of the VAST Applications of Machine Learning, Research Review international Journal of Multidisciplinary, volume 4 and issue 4 in April 2019.
[15] B Surya Samantha, M Trupthi, U Sairam, A review on using crow search algorithms in solving the problems of constrained optimization, International Journal of Scientific Research in Science and Technology,2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY