Breast Cancer Detection Using Deep Learning and CNN-Based Model

[1] KRISHNA BANAVATHU, [2] ALIKANI VIJAYA DURGA

[1] Dept. of ECE, University College of Engineering, Adikavi Nannaya University, Rajamahendravaram, AP.

[2] Dept. of ECE, University College of Engineering, Adikavi Nannaya University, Rajamahendravaram, AP.

Abstract - Breast cancer is the second most dangerous cancer in the world. Most of the women die due to breast cancer not only in India but everywhere in the world. In 2011, USA stated that one in eight women suffered from cancer. Breast cancer develops due to the abnormal cell division in the breast itself which results in the formation of either *Benign* or *Malignant* cancer. So, it is very important to predict breast cancer at an early stage and by providing proper treatment, many lives can be saved. Breast cancer is the most affected disease present in women worldwide 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the USA during 2016 and 40,450 of women's death is estimated. Despite its proven record as a breast cancer screening tool, mammography remains labor-intensive and has recognized limitations, including low sensitivity in women with dense breast tissue. In the last ten years, Neural Network advances have been applied to Breast Histopathology Images to help radiologists increase their efficiency and accuracy. In this project the aim is to use the current knowledge base on convolution neural networks (CNNs) on Breast Histopathology Images. The project first discusses traditional Computer Assisted Detection (CAD) using machine learning and more recently developed CNN-based model for Breast Histopathology Images.

Keywords - Convolution Neural Networks (CNNs), Histopathology Images, Computer Assisted Detection (CAD), machine learning, CNN-based model.

I. INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ml) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

In this paper the analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. Finally to achieve this, we have used Deep learning classification methods to fit a function that can predict the discrete class of new input.

II. LITERATURE WORK

The earlier case breast cancer is detected by various methods and the higher chances of the patient being treated. Therefore, many early detection or prediction methods are being investigated and used in the fight against breast cancer.

In this paper, the aim was to predict and detect Breast cancer early with non-invasive and painless methods that use data mining algorithms. All the data mining classification algorithms in weka were run and compared against a data set obtained from the measurements of an antenna consisting of frequency bandwidth, the dielectric constant of the antenna's substrate, electric field, and tumor information for breast cancer detection and prediction. Results indicate that bagging, ibk, random committee, random forest, and simple cart algorithms were the most successful algorithms, with over 90% accuracy in detection. This comparative study of several classification algorithms for breast cancer diagnosis using a data set from the measurements of an antenna with a 10-fold cross-validation method provided a perspective into the data mining methods' ability of relative prediction and also it was made with the wekatool.in this paper, the weka data mining tool was applied to an antenna dataset to examine the efficacy of data mining methods in the detection of Breast cancer.

The high accuracy rates of these algorithms suggest that breast cancer tumors can indeed be identified non-invasively, at low cost and without exposing patients to harmful radiation, by using data mining classification algorithms, such as bagging, ibk, random committee, random forest, and simple cart.

III.MACHINE LEARNING TECHNIQUES

A major challenge in data mining and machine learning areas is to build an accurate and computationally efficient classifier for medical applications during this study, by utilizing four main algorithms: support vector classifier, random forest, gradient boosting, naive bayes, cart model, neural network and linear regression algorithm on the Wisconsin breast cancer (original) datasets, we tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity, and specificity to find the best classification accuracy. the support vector reaches the accuracy of 98.23% and outperforming other algorithms. Proposed supervised machine learning algorithms: support vector machine, random forest and naive bayes classifiers to classify the breast cancer.

1) *Evaluation of machine learning models:*

Research and prevention of breast cancer have attracted more concern of researchers in recent years, on the other hand the development of data mining methods provides an effective way to extract more useful information from the complex database, and some prediction, classification, and clustering can be made according to the extracted information. In this study, to explore the relationship between breast cancer and some attributes so that the death probability of breast cancer can be reduced, five different classification models including Decision Tree (DT), Random Forest (rf), Support Vector Machine (SVM), Neural Network (NN) and Logistics Regression (LR) are used for the classification of two different datasets related to breast cancer: Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Breast Cancer Database (WBCAD). Three indicators including prediction accuracy values, f-measure metric, and auc values are used to compare the performance of these five classification models. Comparative experiment analysis shows that the random forest model can achieve better performance and adaptation than the other four methods. The results of the prediction will help to decrease the rate of misdiagnoses and make suitable treatment projects for therapy. Provide a reference for experts to distinguish the nature of Breast Cancer. In this study, there are still some limitations that should be solved in further work. For example, though they're also exist some indices people have not found yet, this study only collects the data of 10 attributes in this experiment. The limited raw data affects the accuracy of results. Also, the rf can be combined with other data mining technologies to obtain more accurate and efficient results in future work

IV.PROBLEM SPECIFICATON

The main aim of proposed is to predict Breast Cancer using Deep Learning Conventional Neural Networks model on Breast Histopathology Images (Breast Cancer (BC) specimens scanned) dataset to predict which type of breast cancer either Benign or Malignant This is an image dataset consists of 198,738 IDC negative and 78,786 IDC positive.

Invasive Ductal Carcinoma (IDC) is the most common form of Breast Cancer.

There are 2 directories consist in this dataset indicates the class where 0 is NONIDC and 1 is IDC.

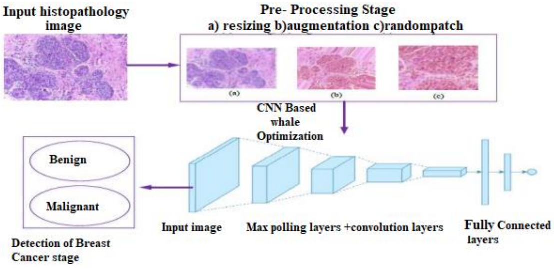

Input histopathology image
Pre- Processing Stage
a) resizing b)augmentation c)randompatch
(a)
(b)
(c)
CNN Based whale Optimization
Benign
Malignant
Input image
Max polling layers +convolution layers
Fully Connected layers
Detection of Breast Cancer stage

Fig 1: Architecture.

Using deep learning CNN algorithm to predict which kind of cancer it is. Giving the highest accuracy, taking images as input and predicts cancer and also give output which kind of cancer it is either benign or malignant.

Prediction of breast cancer is made by the classifier on the test data set after the model is built in the training phase. Prediction is made by 8 attributes given as input and prediction is made for anyone remaining attribute.

1) *Breast Histopathology Images:*

Breast cancer is the most common form of cancer in women, and Invasive ductal Carcinoma (IDC) is the most common form of breast cancer. Accurately Identifying and categorizing breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce error. Invasive Ductal Carcinoma (IDC) is the most common subtype of all breast cancers. To assign an Aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. This dataset contains 277,524 images patches of size 50 x 50 were extracted 198,738 IDC negative and 78,786 IDC positive, the class where 0 is NON-IDC and 1 is IDC.

IDC images

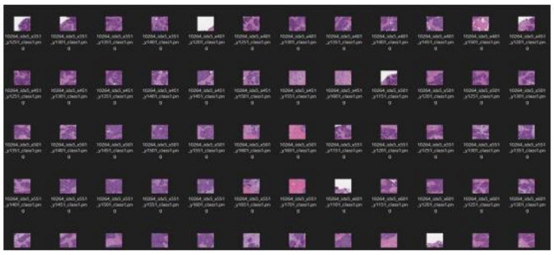


Fig2: IDC Images

NON – IDC images

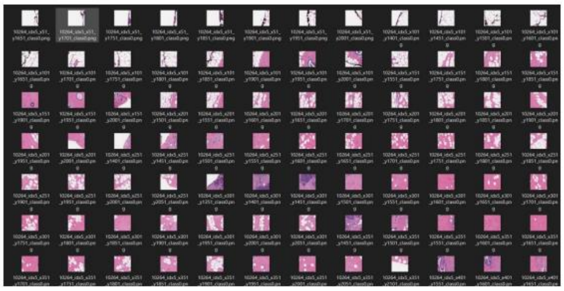


Fig 3:NON-IDC Images

2) *Data Pre-Processing:*

Data pre-processing, data mining technique which is used to transform the raw data in a useful and efficient format. It is the step before applying machine learning algorithms. It transforms the original data into a suitable shape to be used by a particular algorithm. Data pre-processing includes different tasks as data cleaning, feature selection, and data transformation

3) *Data Cleaning:*

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted and the goal of data cleaning is to create data sets that are standardized and uniform to allow business intelligence and data analytics tools to easily access and find the right data for each query. Regardless of the type of analysis or data visualizations you need, data cleaning is a vital step to ensure that the answers you generate are accurate.

4) *Data Transformation:*

Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system

5) *Data Visualization:*

Data visualization is a graphical representation of information and data by using visual elements like charts, graphs and we can quickly identify red from blue, square from the circle. That grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers.

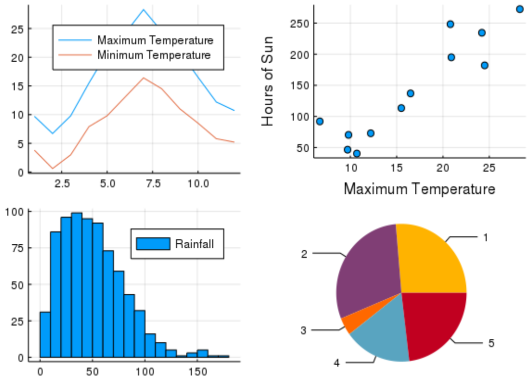
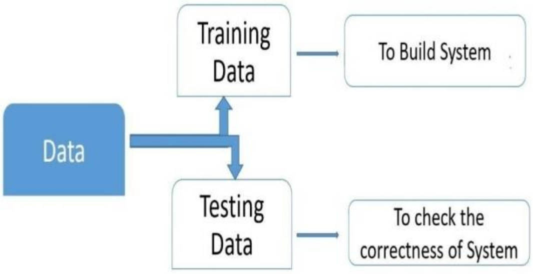


Fig. 4 Data Visualization

DATASET SPLITTING:

It is standard in ml to split data into training and test sets, if you try and evaluate your system on data you have trained it on, separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.



Analysis services randomly sample the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

Train set: The observations in the training set to form the experience that the algorithm uses to learn. in supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.

Test set: Test set is a set of observations used to evaluate the performance of the model using some performance metrics. no observations from the training set must be included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training setor has simply memorized it.

CONVOLUTIONAL NEURAL NETWORKS

Neural networks are typically organized in layers. Layers are made up of several interconnected 'nodes' which contain an 'activation function'. In traditional feed-forward neural networks, each neuron in the input layer is connected to every output neuron in the next layer – we call this a fully connected (fc) layer, however, on CNN, we don’t use fc layers until the very last layers in the network. We can thus define a CNN as a neural network that swaps in a specialized "convolutional" layer in place of a "fully-connected" layer for at least one of the layers in the network.

A nonlinear activation function, such as RELu, is then applied to the output of these convolutions and the process of convolution, activation continues along with a mixture of other layer types to help reduce the width and height of the input volume and help reduce the width and height of the input volume and help reduce over fitting until we finally reach the end of the network and apply one or two fc layers where we can obtain our final output classifications. Each layer in a CNN applies a different set of filters, typically hundreds or thousands of them, and combines the results, feeding the output into the next layer in the network.

During training, a CNN automatically learns the values for these filters. In the context of image classification, our CNN may learn to:

- Detect edges from raw pixel data in the first layer.
- Use these edges to detect shapes in the second layer.
- Use these shapes to detect higher-level features in the highest layers of the network.

The last layer on CNN uses these higher-level features to make predictions regarding the contents of the image in practice, cans give us two key benefits: local invariance and compositionality. The concept of local invariance allows us to classify an image as containing a particular object regardless of where in the image the object appears.

There are many types of layers used to build convolutional neural networks, but the ones we are most likely to encounter include:

- Convolutional layer
- Pooling layer
- Fully-connected layer
- Dropout layer

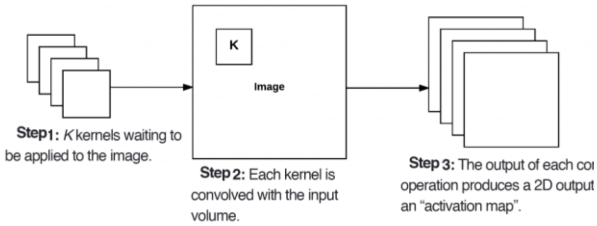

K
Image
Step1: K kernels waiting to be applied to the image.
Step 2: Each kernel is convolved with the input volume.
Step 3: The output of each co
operation produces a 2D output
an "activation map".

Fig 5. Convolution Operation

After applying all k filters to the input volume, we now have k, 2-dimensional activation maps. We then stack our k activation maps along the depth dimension of our array to form the final output volume.

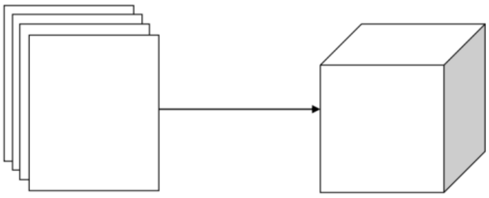


Fig 6 : After obtaining the k activation maps, they are stacked together to form the input volume to the next layer in the network.

A filter or kernel is an integral component of the layered architecture. It is a smallersized matrix in comparison to the dimensions of the image, that contains real-valued entries. The kernels are then convolved with the input volume to obtain so called activation maps. Activation maps indicate activated regions, i.e., regions where features specific to the kernel have been detected in the input.

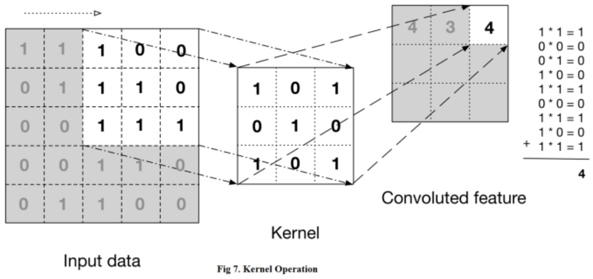


Fig 7. Kernel Operation

Kernel for the different things like image identity, edge detection, sharpening the images.

Stride

We described a convolution operation as "sliding" a small matrix across a large matrix, stopping at each coordinate, computing an element-wise multiplication and sum, then storing the output. This description is similar to a sliding window that slides from left-to-right and top-to-bottom across an image.

The primary function of the pool layer is to progressively reduce the spatial size (i.e., width and height) of the input volume, Doing this allows us to reduce the number of parameters and computation in the network- pooling also helps us control over fitting.

Pool layers operate on each of the depth slices of an input independently using either the max or average function. Max pooling is typically done in the middle of the CNN architecture to reduce the spatial size, whereas average pooling is normally used as the final layer of the network where we wish to avoid using fc layers entirely. Typically, we'll use a pool size of 2x2, although deeper CNNs that use larger input images (>200 pixels) may use a 3x3 pool size early in the network architecture. We also commonly set the stride to either s1 or s-2.

Applying the pool operation yields an output volume of size woutput x houtput x doutput, where:

Woutput = ((winput-f) =s) + 1

Houtput = ((hinput-f) =s) +1

Doutput = dinput

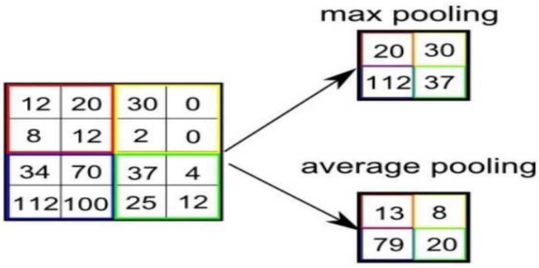


Fig 8: Pool Types

RESULTS

Importing Data Images:

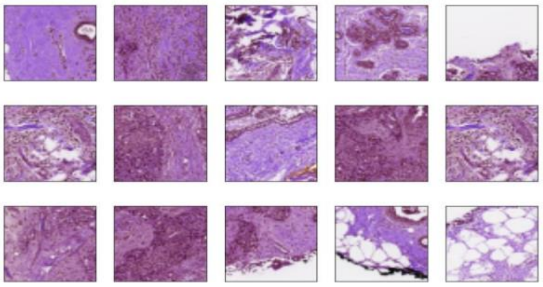
```
data = './10264'
No_breast_cancer = './10264/No'
Yes_breast_cancer = './10264/Yes'
```

```
dirlist=[No_breast_cancer, Yes_breast_cancer]
classes=['No', 'Yes']
filepaths=[]
labels=[]
for i,j in zip(dirlist, classes):
    filelist=os.listdir(i)
    for f in filelist:
        filepath=os.path.join (i,f)
        filepaths.append(filepath)
        labels.append(j)
print ('filepaths: ', len(filepaths), '   labels: ', len(labels))
```

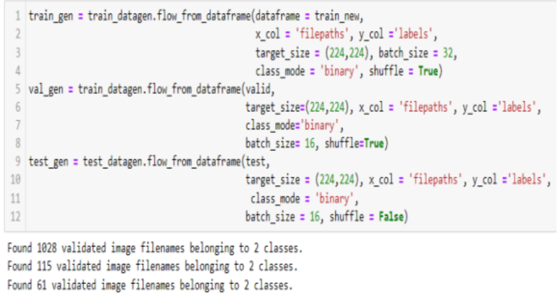
```
filepaths:  1204    labels:  1204
```

Plotting Data Images:

```
#visualize breast tumor images

plt.figure(figsize=(12,8))
for i in range(15):
    random = np.random.randint(1,len(df))
    plt.subplot(3,5,i+1)
    plt.imshow(cv2.imread(df.loc[random,"filepaths"]))
    plt.title(df.loc[random, "labels"], size = 15, color = "white")
    plt.xticks([])
    plt.yticks([])

plt.show()
```



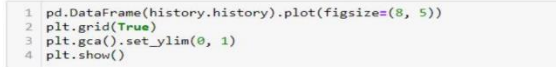
Splitting Training and Testing Data:



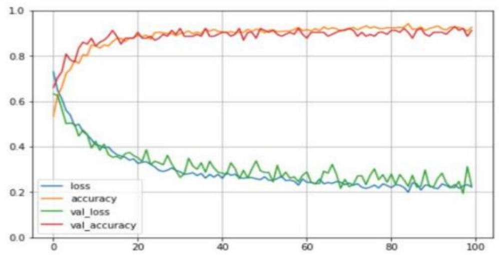
```
train_gen = train_datagen.flow_from_dataframe(dataframe = train_new,
                                              x_col = 'filepaths', y_col ='labels',
                                              target_size = (224,224), batch_size = 32,
                                              class_mode = 'binary', shuffle = True)
val_gen = train_datagen.flow_from_dataframe(valid,
                                            target_size=(224,224), x_col = 'filepaths', y_col ='labels',
                                            class_mode='binary',
                                            batch_size= 16, shuffle=True)
test_gen = test_datagen.flow_from_dataframe(test,
                                            target_size = (224,224), x_col = 'filepaths', y_col ='labels',
                                             class_mode = 'binary',
                                            batch_size = 16, shuffle = False)
```

```
Found 1028 validated image filenames belonging to 2 classes.
Found 115 validated image filenames belonging to 2 classes.
Found 61 validated image filenames belonging to 2 classes.
```

Accuracy Of Data And Loss Of The Data In Graph View:



```
pd.DataFrame(history.history).plot(figsize=(8, 5))
plt.grid(True)
plt.gca().set_ylim(0, 1)
plt.show()
```



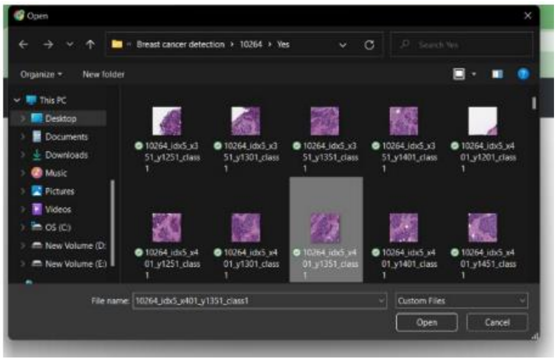
Individual Prediction

:

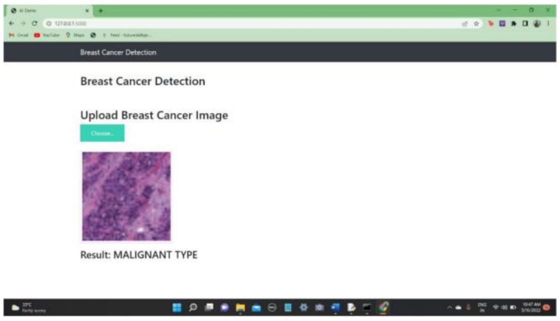
```
from PIL import Image
model_path = "model.h5"
loaded_model = tf.keras.models.load_model(model_path)

# import matplotlib.pyplot as plt
import numpy as np

image = cv2.imread("./10264/No/10264_idx5_x1001_y551_class0.png")

image_fromarray = Image.fromarray(image, 'RGB')
resize_image = image_fromarray.resize((224, 224))
expand_input = np.expand_dims(resize_image,axis=0)
input_data = np.array(expand_input)
input_data = input_data/255

pred = loaded_model.predict(input_data)
if pred >= 0.5:
  print("MALIGNANT")
else:
  print("BENIGN")
```

```
BENIGN
```

Choose an image to predict the cancer:


Open
Breast cancer detection › 10264 › Yes
Organize
New folder
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
New Volume (D:
New Volume (E:)
10264_idx5_x351_y1251_class1
10264_idx5_x351_y1301_class1
10264_idx5_x351_y1351_class1
10264_idx5_x351_y1401_class1
10264_idx5_x401_y1201_class1
10264_idx5_x401_y1251_class1
10264_idx5_x401_y1301_class1
10264_idx5_x401_y1351_class1
10264_idx5_x401_y1401_class1
10264_idx5_x401_y1451_class1
Custom Files
Open
Cancel

File name: 10264_idx5_x401_y1351_class1



Breast Cancer Detection

Upload Breast Cancer Image

Result: MALIGNANT TYPE

CONCLUSION

In this project, we study how to classify breast cancer with deep learning classification algorithms. We propose CNN algorithms to predicting the cancer images. Here the data set taken is Breast Histopathology Images. The dataset is loaded and initially, Exploratory Data Analysis is performed. We perform CNN model building to predict the breast cancer images and derive accuracy. The algorithm with more accuracy is chosen and is used to predict the data.

REFERENCES

[1] Yi-Sheng Sun, Zhao, Han-Ping-Zhu," Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.

[2] AlirezaOsarech, BitaShadgar," A Computer-Aided Diagnosis System for Breast Cancer", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[3] MandeepRana, PoojaChandorkar, AlishibaDsouza, “Breast cancer diagnosis and recurrence prediction using machine learning techniques”, International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.

[4] VikasChaurasia, BB Tiwari and Saurabh Pal – “Prediction of benign and malignant breast cancer using data mining techniques”, Journal of Algorithms and Computational Technology

[5] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction Using Data Mining Method, IEEE Conference paper

[6] D.Dubey, S.Kharya, S.Soni and –“Predictive Machine Learning techniques for Breast Cancer Detection”, International Journal of Computer Science and Information Technologies, Vol.4(6),2013,1023-1028.

[7] Nidhi Mishra, NareshKhuriwal.- “Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm”, 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018

[8] Chao-Ying, Joanne, PengKukLida Lee, Gary M. Ingersoll –“An Introduction to Logistic Regression Analysis and Reporting “, September/October 2002 [Vol. 96(No. 1)