BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK

REPORT OF PROJECT SUBMITTED FOR PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

BY

NAME	CLASS ROLL	UNIV ROLL	UNIV REG NO
SOUMYAMOY DAS	CSE/2018/081	11700118036	181170110105
KAUSHIK MAHAJAN	CSE/2018/089	11700118084	181170110057
SAHEB SARKAR	CSE/2018/078	11700118061	181170110080
ANKIT KUMAR JAISWAL	CSE/2018/065	11700118119	181170110022

UNDER THE SUPERVISION OF

Dr. PARAMA BAGCHI

ASSISTANT PROFESSOR

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA

AT

RCC INSTITUTE OF INFORMATION TECHNOLOGY

[Affiliated to Maulana Abul Kalam Azad University of Technology]

CANAL SOUTH ROAD, BELIAGHATA, KOLKATA – 700015

JUNE – 2022

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA – 700015, INDIA

TO WHOM IT MAY CONCERN

The report of the Project titled **BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK** submitted by SOUMYAMOY DAS (Roll No.: 11700118036), KAUSHIK MAHAJAN (Roll No.: 11700118084), SAHEB SARKAR (Roll No.: 11700118061) and ANKIT KUMAR JAISWAL (Roll No.: 11700118119) of B. Tech. (CSE) 8th Semester of 2022 has been prepared under our supervision for the partial fulfilment of the requirements for Bachelor of Technology in Computer Science and Engineering degree in Maulana Abul Kalam Azad University of Technology. The report is hereby forwarded.

_____

Dr. Parama Bagchi

Dept. of Computer Science and Engineering

RCCIIT, Kolkata – 700015

Countersigned by:

_____

Mr. Rajib Saha

Head of the Department
Computer Science and Engineering
RCCIIT, Kolkata – 700015

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA – 700015, INDIA

CERTIFICATE OF APPROVAL

The report of the Project Titled BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK submitted by SOUMYAMOY DAS (Roll No.: 11700118036) of B. Tech. CSE is hereby recommended to be accepted for the partial fulfilment of the requirements for B. Tech. (CSE) degree in Maulana Abul Kalam Azad University of Technology.

Name of Examiner	**Signature with Date**
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA – 700015, INDIA

CERTIFICATE OF APPROVAL

The report of the Project Titled BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK submitted by KAUSHIK MAHAJAN (Roll No.: 11700118084) of B. Tech. CSE is hereby recommended to be accepted for the partial fulfilment of the requirements for B. Tech. (CSE) degree in Maulana Abul Kalam Azad University of Technology.

Name of Examiner	**Signature with Date**
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA – 700015, INDIA

CERTIFICATE OF APPROVAL

The report of the Project Titled BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK submitted by SAHEB SARKAR (Roll No.: 11700118061) of B. Tech. CSE is hereby recommended to be accepted for the partial fulfilment of the requirements for B. Tech. (CSE) degree in Maulana Abul Kalam Azad University of Technology.

Name of Examiner	**Signature with Date**
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____

RCC INSTITUTE OF INFORMATION TECHNOLOGY

KOLKATA – 700015, INDIA

CERTIFICATE OF APPROVAL

The report of the Project Titled BREAST CANCER DIAGNOSIS USING DEEP LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK submitted by ANKIT KUMAR JAISWAL (Roll No.: 11700118119) of B. Tech. CSE is hereby recommended to be accepted for the partial fulfilment of the requirements for B. Tech. (CSE) degree in Maulana Abul Kalam Azad University of Technology.

Name of Examiner	**Signature with Date**
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____

ACKNOWLEDGEMENT

We express our sincere gratitude to Dr. Parama Bagchi of Department of Computer Science and Engineering, RCC Institute of Information Technology and for extending her valuable time for us to take up this problem as a Project.

We are also indebted to Mr. Rajib Saha and Dr. Dipankar Majumdar for their unconditional help and inspiration.

Name: Soumyamoy Das
Roll No.: CSE/2018/081 [11700118036]

Name: Kaushik Mahajan
Roll No.: CSE/2018/089 [11700118084]

Name: Saheb Sarkar
Roll No.: CSE/2018/078 [11700118061]

Name: Ankit Kumar Jaiswal
Roll No.: CSE/2018/065 [11700118119]

Bachelor of Technology
Computer Science and Engineering
8th Semester 2022
RCCIIT Kolkata

Date: 17th June 2022

Abstract

Cancer is the one of the most dangerous diseases in the world. So, our prime target must be curing the cancer through scientific investigation and the second main target should be early detection of cancer because the early detection of cancer can be helpful for curing the cancer completely. After reviewing several papers, we have found that several techniques are available for cancer detection. In this paper, we have proposed Deep Leaning algorithm neural network for diagnosing breast cancer using Wisconsin Breast Cancer database. The paper shows how we can use deep learning technology for diagnosing breast cancer using UCI Dataset. Because deep learning techniques almost used for high task objective Computer Vision, Image processing, Medical Diagnosis, Neural Language Processing. But in this paper, we are applying deep learning technology on the Wisconsin Breast Cancer Database and we have seen that is very beneficial for diagnosing breast cancer. This paper is divided in three parts: first we have collected dataset and applied pre-processing algorithm for scaled and filter data then we have split dataset in training and testing purpose and generate some graph for visualization data. At last, we are implementing the model. So, we have seen deep learning technology is a good way for diagnosis breast cancer with Wisconsin Breast Dataset. This database provides 569 rows and 30 features in the dataset. In this paper we have used 30 features for diagnosing breast cancer that we have got after pre-processing. But before training model, we have applied some pre-processing methods like Label Encoder and Min Max Scaler for scaling our dataset and then used in our model to achieve accuracy.

Table of Contents

List of Symbols

SYMBOL	MEANING	PAGE
$\sum$	Sigma: Summation	24
$\varnothing$	Phi	24

List of Abbreviations

ABBREVIATION	MEANING
CNN	Convolutional Neural Network
DL	Deep Learning
ML	Machine Learning
RELU	Rectified Linear Unit
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
FNAC	Fine Needle Aspirate Cytology
WDBC	Wisconsin Diagnostic Breast Cancer
CSV	Comma Separated Values
OS	Operating System
RAM	Random Access Memory
CPU	Central Processing Unit
GPU	Graphics Processing Unit
IDE	Integrated Development Environment

List of Figures

Chapter 1

Introduction

Cancer is one of the deadliest diseases of the world. The exact cause of cancer is still unknown even today. There is no specific medication to eliminate cancer. The medicines or therapies which are utilised to reduce the rate of cancer are either highly expensive or painful for the patient. Sometimes, those therapies result in very serious side effects. Moreover, cancer is curable only when diagnosed at an early stage, else it proves fatal to many.

There are various forms of cancer like blood cancer, lung cancer, ovarian cancer, breast cancer among others. For females, breast cancer is one of the most common in the world. The life time risk of developing a breast cancer in women is approximately 1 in 8 in the USA, 1 in 12 in Europe and 1 in 40 in Asia (though this is the old statistics of WHO 2008). The number is increasing every year.

In India, according to cancerconsultindia.com, there are more than 1.38 million new breast cancer cases every year. Which means, every 4 minutes, an Indian woman is diagnosed with breast cancer. Moreover, there are more than 4 lakhs fatalities from breast cancer every year in India, which means one woman falls prey of breast cancer every 13 minutes.

Hence, it is of utmost importance to diagnose and treat breast cancer at the earliest stage so as to save these innocent lives. However, cancer detection is mainly done using Biopsy or Fine Needle Aspirate Cytology (FNAC) examination, where the breast tumorous tissue is processed and then studied manually under a microscope. This process is very much time consuming and generally the results are declared after 2 to 3 days. Although, there are other techniques like mammography to detect cancer, but the accuracy rate of mammography is lesser than FNAC and also the setup is quite costly.

Deep Learning: In a word, accuracy. Deep learning achieves recognition accuracy at higher levels than ever before. This helps consumer electronics meet user expectations, and it is crucial for safety-critical applications like driverless cars. Recent advances in deep learning have improved to the point where deep learning outperforms humans in some tasks like classifying objects in images.

While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

Deep learning requires large amounts of labeled data. For example, driverless car development requires millions of images and thousands of hours of video.

Deep learning requires substantial computing power. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

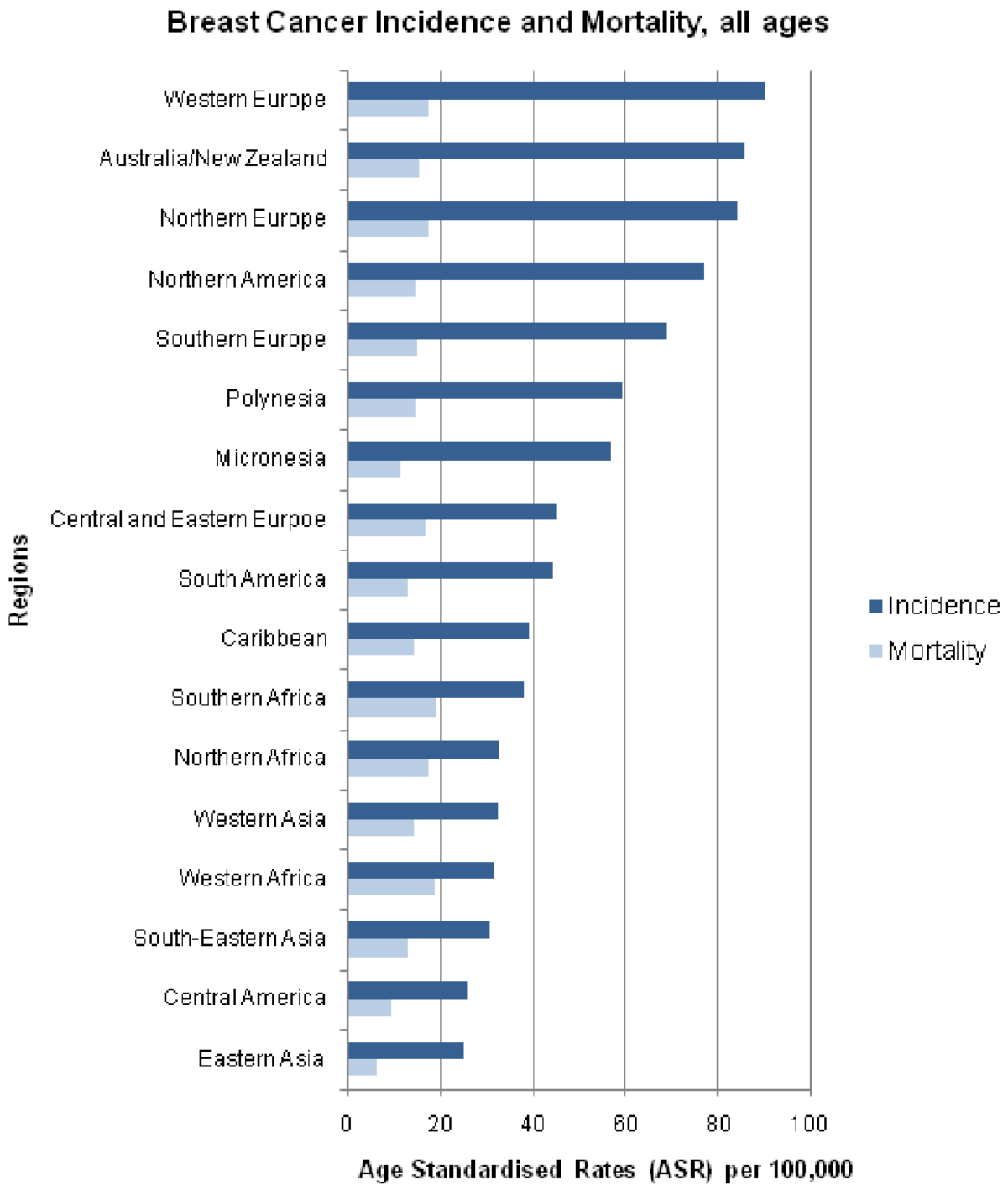

Figure 1: Breast Cancer Incidents and Mortality (Globally)

Figure 2: Cancer Statistics in India

Literature Review

Sl. No.	Author Names	Paper Name	Contribution
01	Md. Abu Ismail Siddique, Md. Omaer Faruq Goni, Md. Habibur Rahaman, Fahim Md. Sifnatul Hasnain and Oishi Jyoti	Breast Cancer Detection using Deep Neural Network	Breast Cancer Detection using the WDBC Dataset
02	Naresh Khuriwal, Nidhi Mishra	Breast Cancer Diagnosis Using Deep Learning Algorithm	Breast Cancer Detection using the WDBC Dataset
03	Ram Murti Rawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal	Breast Cancer Detection Using K-Nearest Neighbours, Logistic Regression and Ensemble Learning	Breast Cancer Detection using the WDBC Dataset
04	Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, Gunter Saake	Analysis of Breast Cancer Detection Using Different Machine Learning Tech-niques	Breast Cancer Detection using the WDBC Dataset

Chapter 2

Workflow

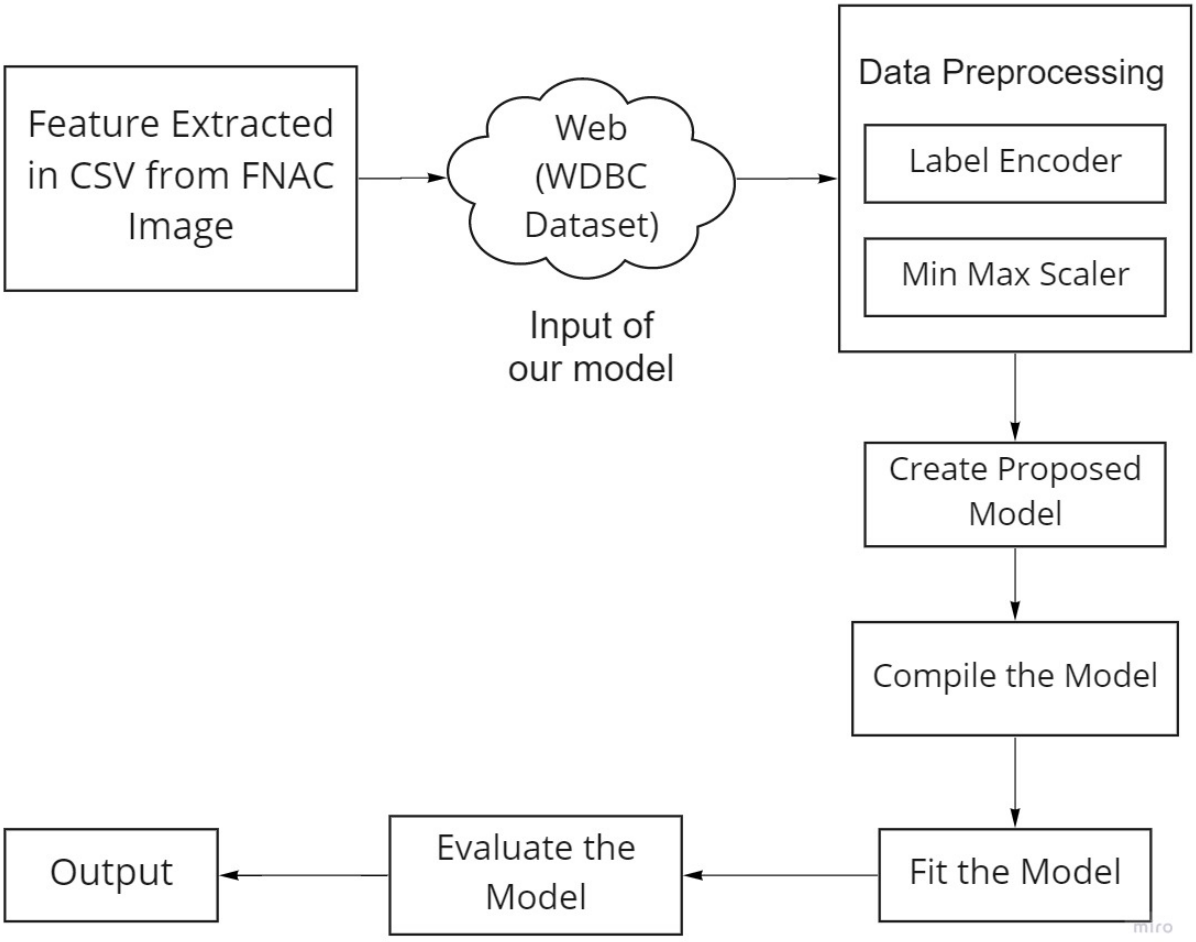

Figure 3: Workflow

Methodologies of Implementation

Analysis

Our project aims to speed up the process for detection of breast cancer, eliminating manual examination of tissues under microscope.

We have tried to create a Deep Learning model based on Convolutional Neural Network (CNN) to predict breast cancer once the features of the cell nuclei are extracted from the FNAC digitized image. To serve this purpose, we are using the Wisconsin Diagnostic Breast Cancer (WDBC) Data Set to train and test our model. The data set is owned and maintained by the University of Wisconsin, Clinical Studies and can be found here from the UCI Machine Learning Repository.

The data set comprises of 569 unique entries (CSV) of cancerous and non-cancerous cells. There are 32 attributes comprising of 10 features, extracted from sample FNAC images. Every feature is presented in 3 forms: mean, standard error and worst. Hence, we are using these 30 attributes to train and test our model.

Ten real-valued features computed for each cell nucleus are:

a) radius (mean of distances from centre to points on the perimeter)

b) texture (standard deviation of grey-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

Among those 569 entries, 357 are Benign and 212 are Malignant entries.

This problem is basically a 2-class problem:

- B-Benign
- M-Malignant

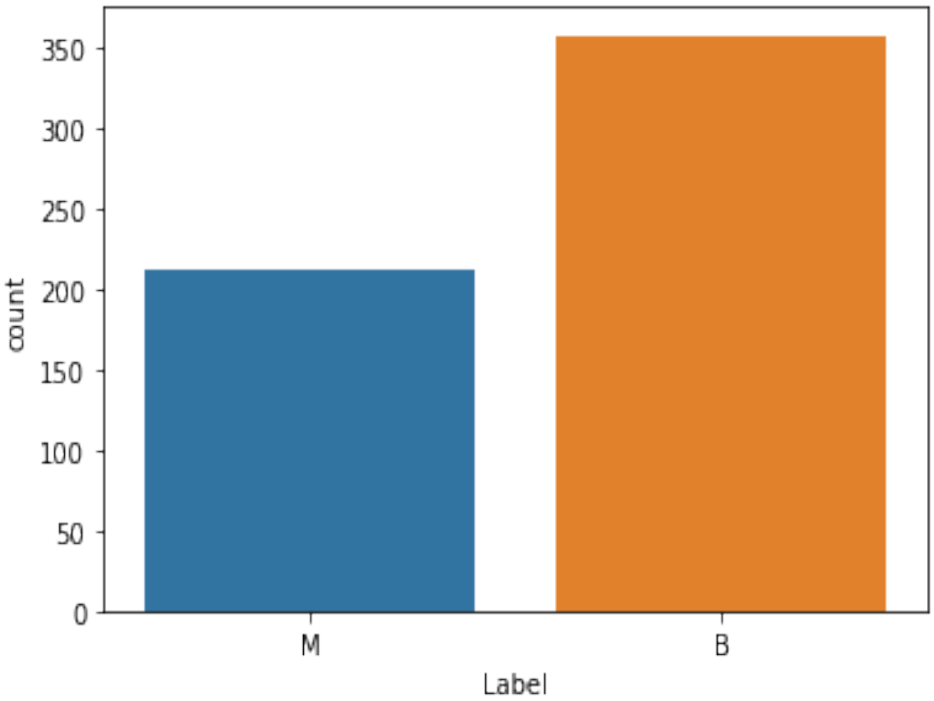

Figure 4: Number of Malignant and Benign entries

Algorithm

1. So far, we have analysed that this problem is a 2-class problem, i.e., a Binary Classification problem. *We have trained a deep learning model using* ***Convolutional Neural Network (CNN)****. The input layer of our model will comprise of 30 neurons, representing 30 unique features.* The output layer will basically be 0 or 1, according to the prediction. In between we can add several hidden layers to improve our model. We are using 75% of input data for training purpose and the remaining for testing.

2. Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It's achieving results that were not possible before.

3. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labelled data and neural network architectures that contain many layers.

4. One of the most important steps in deep learning is data pre-processing. Data pre-processing is a data mining technique that used to filter data in a usable format. Since the real-world dataset is available in different formats, so it must be filtered to fit it into our model.

5. Label Encoding:

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

6. Normalization:

Normalization methods allow the transformation of any element of an equivalence class of shapes under a group of geometric transforms into a specific one, fixed once for all in each class.

We are using Python to create the model, and thus various libraries of Python like pandas, numpy, sklearn, scikit-learn, seaborn, tensorflow, matplotlib, keras could be used.

7. Prediction: The predicted value always ranges from 0.0 to 1.0. So, the value <= 0.5 is taken as 0 and the value > 0.5 is taken as 1 (where 0 means Malignant and 1 is Benign).

8. After the prediction, we are using a confusion matrix to summarize the results and is also used for evaluating the performance of the classification model.

9. Finally, we have to calculate the accuracy from the confusion matrix. Accuracy is the number of correct predictions made by the model over all kind of predictions made. We can find the accuracy using the below formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where, TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Implementation Details

1. Importing libraries and data set: After importing all the necessary Python libraries in our code, we are reading the CSV file using pandas, and put in data frame.

2. Data Preprocessing:

In the data pre-processing part, we have to check for missing data or `null` values and remove them, if any. In the Wisconsin data set, there are no `null` values. Hence the data is an excellent data, i.e., it is a clean data.

Next, we are renaming the Diagnosis column in dataset to 'Label'. Then we can define the dependent variables that needs to be predicted (labels).

3. Label Encoding:

Now we can then do encoding of categorical data from text (B and M) to integers(0 and 1). This will be done using LabelEncoder() method from sklearn.preprocessing.

sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values. LabelEncoder encode labels with a value between 0 and n_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier. This transformer is used to encode target values, and not the input.

4. Then we have to define the independent variables.

5. MinMax Scaler:

Now we normalize/scale the values to bring them to similar range. This is being done using MinMaxScaler from sklearn.preprocessing. Feature scaling in machine learning/deep learning is one of the most critical steps during the pre-processing of data before creating a machine learning/deep learning model. Scaling can make a difference between a weak machine learning model and a better one. In many algorithms, when we desire faster convergence, scaling is a must like in Neural Network. Min-Max Scaler transform features by scaling each feature to a given range. This estimator scales and translates each feature

individually such that it is in the given range on the training set, e.g. between 0 and 1.

The min-max transformation is given by:

$$x_{std} = \frac{x - min(x)}{max(x) - min(x)}$$

$$x_{scaled} = x_{std} * (max - min) + min$$

where `min - max = feature range`

This transformation is often used as an alternative to zero mean, unit variance scaling.

6. Dataset splitting:

Next, we have to split the data into training data and testing data, to verify accuracy after fitting the model. This splitting can be done using train_test_split from sklearn.model_selection.

The train_test_split function is for splitting a single dataset for two different purposes: training and testing. The training subset is used for building the model. The testing subset is for using the model on unknown data to evaluate the performance of the model. With this function, we don't need to divide the dataset manually.

In our data set, there are 569 datapoints. Our test size is 25%, thus Training data is (426,30) and Testing data is (143,30).

7. Building the model:

Now comes the deep learning part. We will make a deep learning model based on Convolutional Neural Network (CNN) to detect whether the cells are cancerous or non-cancerous.

Sequential Model:

Sequential is the easiest way to build a model in Keras. It allows us to build a model layer by layer. We use the add() function to add layers to our model and

dropout() function to specify the percentage of connections to be dropped before moving to the next layer. We are using sequential method, it makes adding the layers easy.

In our neural network model, we have used a total of 5 layers, comprising of 1 input layer, 3 hidden layers and another 1 output layer.

The first input layer comprises of 30 neurons representing the 30 unique input features. The next hidden layer comprises of 20 neurons. The third one comprises of 12 neurons, and the fourth one 4 neurons. The final one comprises of one and only neuron, representing the final result, 0 or 1.

Activation Functions:

Rectified Linear Unit:

The activation function used in all the dense layers except the output one, is *relu*. Relu stands for *Rectified Linear Unit*. Mathematically, it is defined as $y = max(0, x)$. Relu is the most commonly used activation function in neural networks, especially in CNNs. It is a linear function that will output the input directly if it positive, otherwise it will output zero. We use this function because a model that uses it, often achieves better performance. The rectified linear activation is the default activation when developing multilayer Perceptron and convolutional neural networks. It converges faster. Linearity means that the slope doesn't plateau, or saturate, when x gets large. It doesn't have the vanishing gradient problem suffered by other activation functions like sigmoid or tanh.

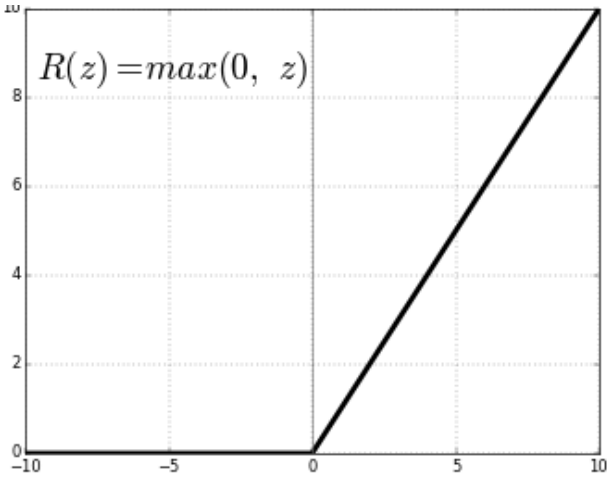

Figure 5: Rectified Linear Unit function

Sigmoid:

The activation function used in the output layer is *sigmoid*. The sigmoid activation function (also known as *logistic function*), is a very popular activation function for neural networks. The *sigmoid* function curve looks like a *S-shape*. The sigmoid activation function is used here because the input to the function is transformed into a value between 0.0 and 1.0.The inputs that are larger than 1.0 are transformed to the value 1.0 and the inputs that are smaller than 0.0 are transformed into value 0.0. The function is differentiable, which means, we can find the slope of the sigmoid curve at any two points. The function is monotonic but function's derivative is not.

$$\varnothing(z)=\frac{1}{1+e^{-z}}$$

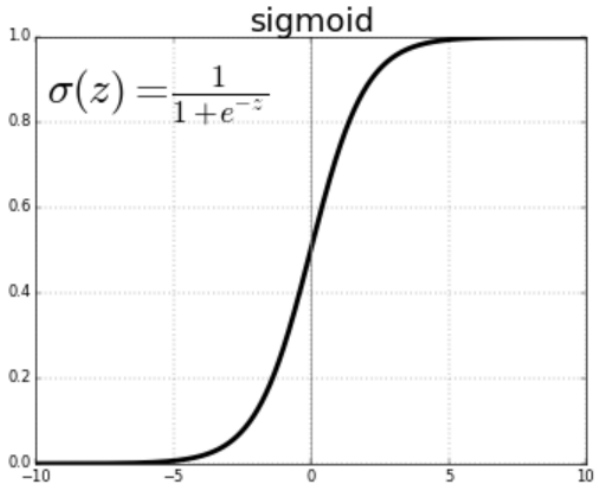

Figure 6: Sigmoid function

Overfitting may occur when large neural nets are trained on relatively small datasets. The performance of the model on new data is negatively impacted when overfitting occurs. This also negatively impacts the model's ability to generalize. Dropout is a simple way to prevent neural networks from overfitting. Hence, we are dropping out 10% of the connections and every layers.

During training, some number of layer outputs are randomly ignored or dropped out. Due to this, the layer is treated like a layer with a different number of nodes and connected to the prior layer.

Binary Cross Entropy:

The model is then compiled using *binary cross entropy* loss function.

Loss/cost functions are used to optimize the model during training. The objective is almost always to minimize the loss function. The lower the loss the better the model. Cross-Entropy loss is a most important cost function. It is used to optimize classification models.

The loss function is used to compute the quantity that the model should seek to minimize during training. The mean squared error is the commonly used loss function for regression models, while for classification models predicting the probability, cross entropy is the most commonly used loss function.

Binary_crossentropy is one of the three different types of cross entropy loss functions, provided by Keras. It is used as a loss function for binary classification model. It computes the cross-entropy loss between true and predicted labels.

Neural Network Design

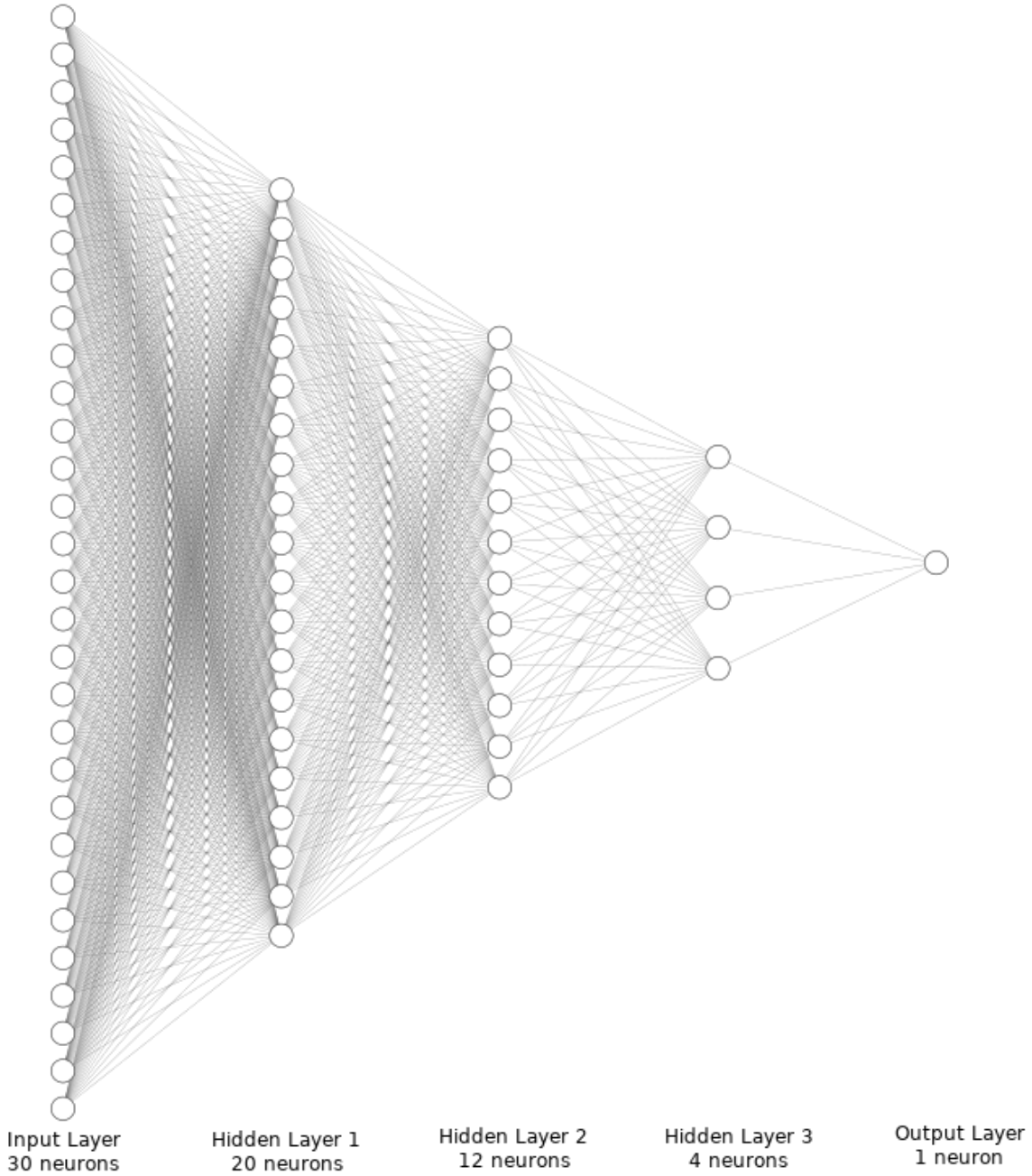

Figure 7: Neural Network of our Model

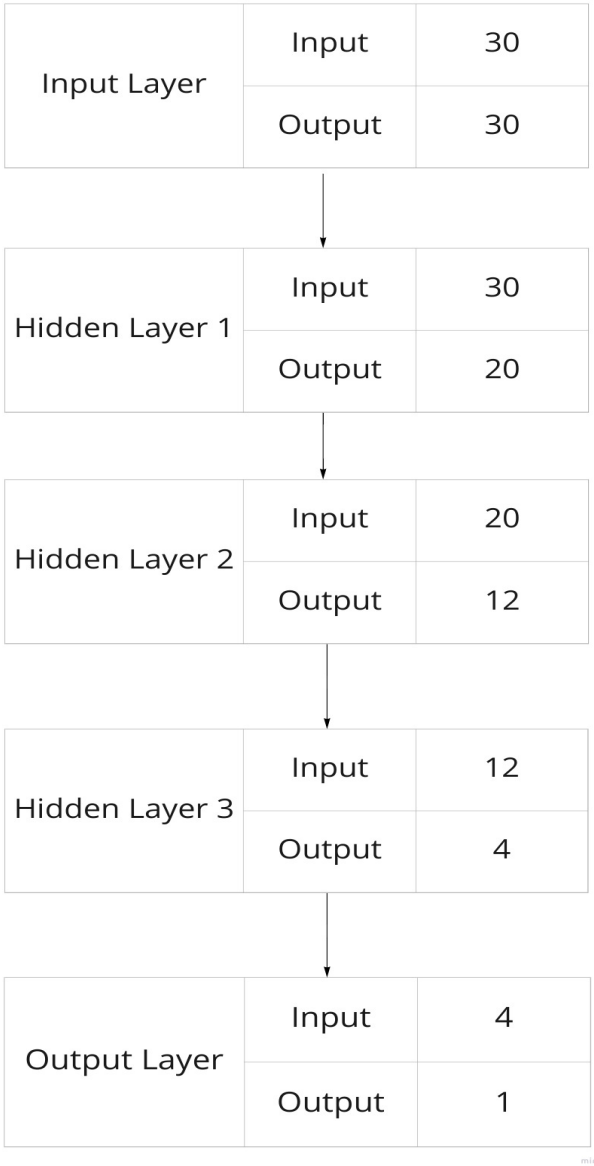

Figure 8: Deep Learning Model (based on CNN) Diagramatic Representation

Software and Hardware Requirements

Software Specifications

We have used Python as the Programming Language to implement the model.

Since we are having hardware constraints, so we are using Google Colaboratory as the environment/IDE/compiler to compile our model.

If this model is to be compiled and executed locally, then minimum software specifications required would be:

- Operating System: Windows/MacOS/any Unix – family OS
- Compiler: Python (preferable latest version)
- Dependencies/Libraries: pandas, numpy, sklearn, scikit, keras, seaborn, matplotlib

Hardware Specifications

Since we have used Google Colaboratory, so no specific hardware is required in our local machines.

However, for compiling the model in local machine, preferred hardware specifications would be:

- Primary Memory (RAM): minimum 4 GB
- Main Processor (CPU): Any multicore CPU
- Graphics Processing Unit (GPU): Preferred for better execution experience
- Secondary Memory: minimum 100 MB of free secondary memory

Chapter 3

Test Cases and System Validation

The dataset we are using is the Wisconsin Breast Cancer Dataset by the University of California Irvine. It comprises of 569 unique entries and 32 attributes. Out of the 569 entries, we are using 75% of the data for training purpose and the rest 25% for testing.

The number of epochs used while training the model was 100 and batch size was 64.

Table 1: Execution of last 20 epochs

Epoch	Loss	Accuracy	Value of Loss	Value of accuracy
81	0.1429	0.9601	0.0716	0.9650
82	0.1342	0.9507	0.0714	0.9720
83	0.1226	0.9718	0.0715	0.9790
84	0.1652	0.9624	0.0695	0.9650
85	0.1421	0.9507	0.0686	0.9720
86	0.1416	0.9601	0.0715	0.9790
87	0.1439	0.9601	0.0685	0.9720
88	0.1238	0.9531	0.0699	0.9650
89	0.1311	0.9671	0.0670	0.9720
90	0.1366	0.9624	0.0680	0.9790
91	0.1269	0.9554	0.0672	0.9790
92	0.1350	0.9648	0.0658	0.9790
93	0.1137	0.9671	0.0665	0.9790
94	0.1137	0.9671	0.0665	0.9790

95	0.1046	0.9671	0.0651	0.9790
96	0.1024	0.9742	0.0739	0.9720
97	0.1338	0.9577	0.0661	0.9790
98	0.1192	0.9671	0.0653	0.9790
99	0.1180	0.0954	0.0690	0.9720
100	0.1230	0.9671	0.0641	0.9790

Observed Output

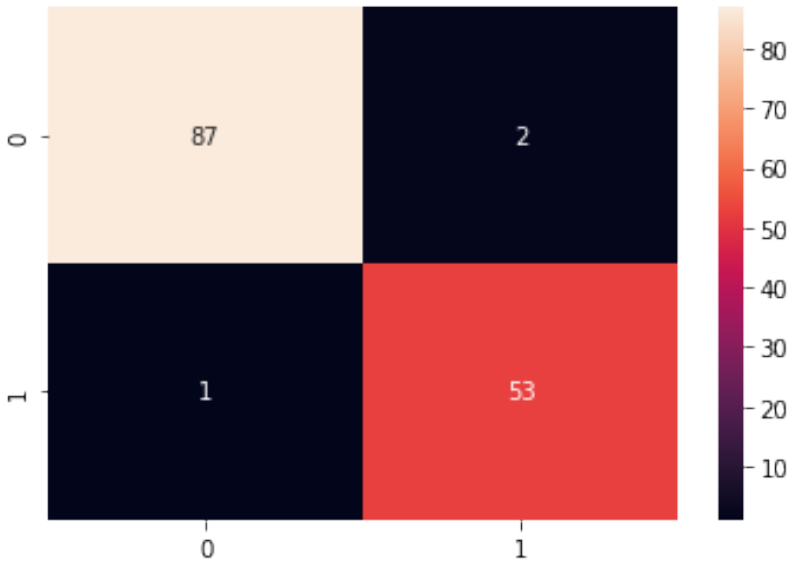

Figure 9: Confusion Matrix of our model

Finally, we have to calculate the accuracy from the confusion matrix. Accuracy is the number of correct predictions made by the model over all kind of predictions made. We can find the accuracy using the below formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where, TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Therefore, overall accuracy of our model from the Confusion matrix is 0.97902.

Performance Analysis

Training Loss:

The training loss is a metric used to assess how a deep learning model fits the training data. That is to say, it assesses the error of the model on the training set. Note that, the training set is a portion of a dataset used to initially train the model. Computationally, the training loss is calculated by taking the sum of errors for each example in the training set.

It is also important to note that the training loss is measured after each batch. This is usually visualized by plotting a curve of the training loss.

Validation Loss:

Validation loss is a metric used to assess the performance of a deep learning model on the validation set. The validation set is a portion of the dataset set aside to validate the performance of the model. The validation loss is similar to the training loss and is calculated from a sum of the errors for each example in the validation set.

Additionally, the validation loss is measured after each epoch. This informs us as to whether the model needs further tuning or adjustments or not. To do this, we usually plot a learning curve for the validation loss

Below we have plotted both the metrics:

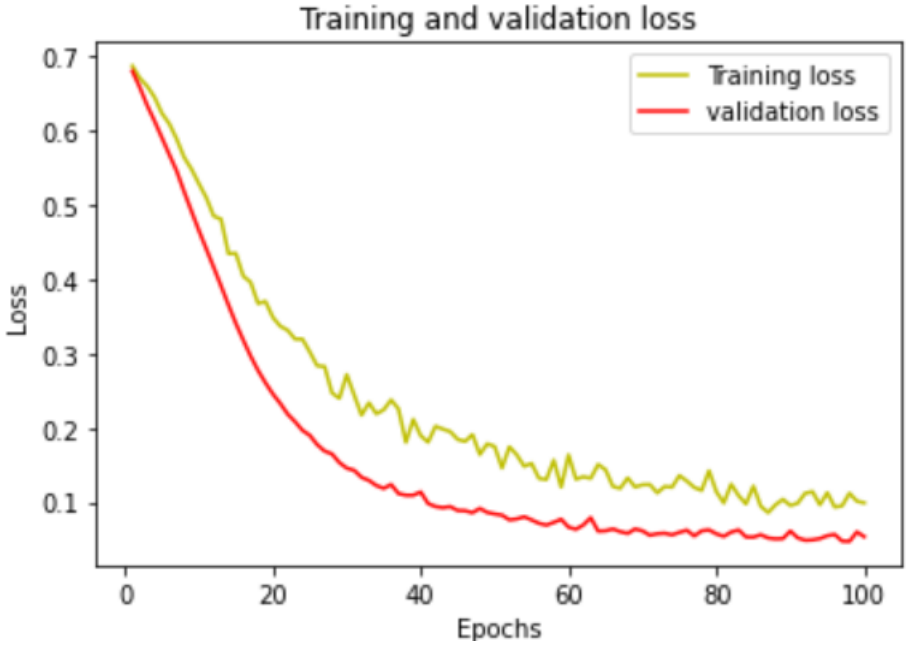

Figure 10: Training vs Validation loss curve

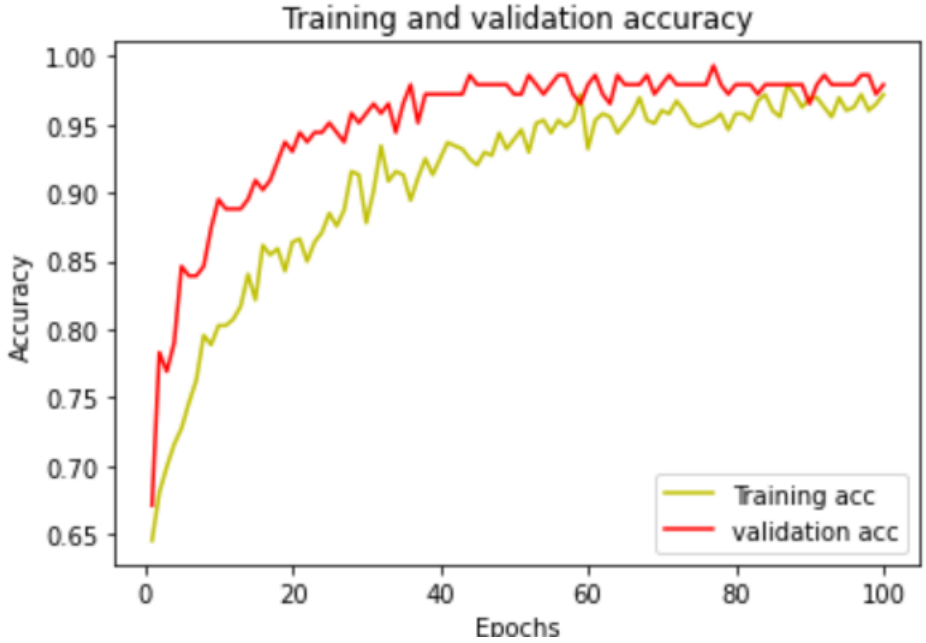

Figure 11: Training vs Validation accuracy curve

Chapter 4

Conclusion

Through our model, we have been able to achieve accuracy of over 97%. This procedure can be adopted by the diagnostic centres to predict whether the given tissue is cancerous or not, since our precedure is accurate enough. It would eliminate any manual examination of the FNAC tissues under a microscope and would save a lot of time and prevent human errors. This might would also cut down the cost of tests, and might make it more affordable for the normal people. This project could have been developed further if we can get more support in terms of ideas and technology.

Future Scope

We have considered the WDBC Data Set as input to our model and we are implementing deep learning model based on CNN directly on the features pre-extracted from the FNAC images. Therefore, for our model to predict, the features should be directly provided as input in the form of CSV. Thus, some other algorithm is required to extract the features from the FNAC images. If we could have directly extracted the features from the images and implement the Machine Learning model, it would have been better and more concise.

References

1 WDBC Data Set (UCI ML Repository): https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

2 TensorFlow Documentation: https://www.tensorflow.org/api_docs

3 Breast Cancer Statistics: Google and https://cancerconsultindia.com/blog/breast-cancer-statistics-rise-of-breast-cancer-in-india/

4 Breast Cancer Wikipedia: https://en.wikipedia.org/wiki/Breast_cancer

5 Google Colaboratory: https://colab.research.google.com/

6 Md. Abu Ismail Siddique, Md. Omaer Faruq Goni, Md. Habibur Rahaman, Fahim Md. Sifnatul Hasnain and Oishi Jyoti, 2020 23rd International Conference on Computer and Information Technology (ICCIT), "Breast Cancer Detection using Deep Neural Network"

7 N. Khuriwal, N. Mishra, International Conference on Advances in Computing, Communication Control and Networking (ICACCCN2018), "Breast Cancer Diagnosis Using Deep Learning Algorithm"

8 Ram Murti Rawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), "Breast Cancer Detection Using K-Nearest Neighbours, Logistic Regression and Ensemble Learning"

9 Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, Gunter Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques"

10 Mohamad Mahmoud Al Rahhal, "Breast Cancer Classification in Histopathological Images using Convolutional Neural Network" International Journal of Advanced Computer Science and Applications (IJACSA), 9(3), 2018.

11 M. H. Yap et al., "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," in IEEE Journal of Biomedical and Health Informatics, July 2018.

12 Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. (2017) Classification of breast cancer histology images using Convolutional Neural Networks. PLoS ONE 12(6): e0177544. https://doi.org/10.1371/journal.pone.0177544.

13 H. Song, A. Men and Z. Jiang, "Breast tumor detection using empirical mode decomposition features," in IEEE Access.

14 M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017

15 R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), Kolkata, 2017

16 S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017

17 Y. Yang, P. A. Fasching and V. Tresp, "Predictive Modeling of Therapy Decisions in Metastatic Breast Cancer with Recurrent Neural Network Encoder and Multinomial Hierarchical Regression Decoder," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, 2017

18 Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, 2017

19 A. Alzubaidi, G. Cosma, D. Brown and A. G. Pockley, "Breast Cancer Diagnosis Using a Hybrid Genetic Algorithm for Feature Selection Based on Mutual Information," 2016 International Conference on Interactive Technologies and Games (ITAG), Notthingham, 2016

20 A. I. Pritom, M. A. R. Munshi, S. A. Sabab and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2016

21 A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th International Symposium on Health Informatics and Bioinformatics, Antalya, 2010

22 A. Qasem et al., "Breast cancer mass localization based on machine learning," 2014 IEEE 10th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, 2014

23 B. M. Abed et al., "A hybrid classification algorithm approach for breast cancer diagnosis," 2016 IEEE Industrial Electronics and Applications Conference (IEACon), Kota Kinabalu, 2016

24 B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016

25 C. Deng and M. Perkowski, "A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, 2015

26 D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016

27 D. T. Saleh, A. Attia and O. Shaker, "Studying combined breast cancer biomarkers using machine learning techniques," 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herlany, 2016

28 H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bengaluru, 2013

29 Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong, Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation, Biomedical Research Volume 29, Issue 10, (2018).

30 L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proenc ̧a, "Digital signature of network segment for healthcare environments support," Irbm, vol. 35, no. 6, pp. 299-309, 2014.

31 D. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37-63, 2011.

32 T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," Stat. Comput., vol. 21, no. 2, pp. 137-146, 2011.

33 R. J. Manoj, M. A. Praveena, and K. Vijayakumar, "An aco–ann based feature selection algorithm for big data," Cluster Computing, vol. 22, no. 2, pp. 3953–3960, 2019.

34 O. I. Obaid, M. A. Mohammed, M. Ghani, A. Mostafa, and F. Taha, "Evaluating the performance of machine learning techniques in the classification of wisconsin breast cancer," International Journal of Engineering & Technology, vol. 7, no. 4.36, pp. 160–166, 2018.