Chapter 1: Introduction

Nowadays, there is a drastic level of increase in the amount of women around the world, getting affected by breast cancer and it is raising over time. Earlier detection of breast cancer reduces death rate and avoids it till reaches the chronic level. In 2012, almost 1.7 million women are affected by breast cancer. The impact of breast cancer is increasing day by day, due to that the healthcare professionals are not in a condition to state the people affected at different stages earlier to save their lives. Digital mammography is a major diagnosis model used throughout the world for breast cancer detection. Computer-aided diagnosis (CAD) is widely used in detecting numerous diseases with accurate decision. It assists the healthcare professionals to analyze and conclude the stages of the various diseases. CAD systems are developed in a way to provide promising results and perfect decisions on patient's condition that helps medical practitioners to diagnose the stages of the diseases. It supports radiologists to avoid misconceptions and wrong diagnosis due to inaccurate data, lack of focus, or inexperience, who uses visually screening mammogram of patients. The aim of the system is to develop a novel CAD model to diagnose a breast cancer in earlier stages with more accurate results to save their precious lives by using CNN.

CNN followed by a similar artificial neural network with a complex network structure which has 'n' hidden layers, which can process the input data from the previous layer. The error rate of the input data will be consistently reduced by adjusting the weights of every node, which leads to achieve an accurate result. It helps to create a model and define its complex hierarchies in a simple form. It supports all kinds of algorithms, namely supervised, unsupervised, semi-supervised, and reinforcement. So, the system did not define any specific algorithm. The CNN generates a better model themselves to train the given data.

Chapter 2: Literature Review

Neural Network is inspired by the working principle of biological neural networks, which has its own input and output channels called as dendrites and axons, respectively. A typical ANN will have millions of processing units or elements, which forms a highly interconnected network that processes a huge amount of information based on the response, given by the external input of a computing system. Every single neuron in a typical neural network is called as unit. A layer in a neural network is considered as a set of neurons in a stack. A layer may have n number of nodes in it. A typical neural network system has single input layer and may have a single or two hidden layers, which is directly connected to the output layer, which receives input from the input layer, i.e., previous layer of the current node.

Abbas developed a system with pareto-differential evaluation algorithm with local search scheme, called memetic pareto-artificial neural network (MPANN). MPANN analyzes the data effectively than other models. The method achieved 98.1% accuracy on random split. Tuba and Tulay proposed the statistical neural network-based breast cancer diagnosis system. In the diagnosis system, they used RBF, general regression neural network (GRNN), and statistical neural network structures on WDBC dataset. The system obtained 98.8% on 50–50 partitioning split.

Paulin and Santhakumaran developed a system with back-propagation neural network (BPNN) and obtained 99.28% accuracy with Levenberg–Marquardt algorithm. They used median filter for preprocessing and normalized the data using min–max technique. None of the features are eliminated from the dataset. The accurate result attained from 80:20 partition scheme. Karabatak and Ince developed an expert system for breast cancer detection. Association rules (AR) are used to reduce the dimensions of the dataset. In the system, AR1 and AR2 are developed to reduce the features. AR1 reduces one feature from 9, and AR2 reduces

5 features out of 9. The conventional neural network is used for classification in both AR1 and AR2. The method attained 95.6% accuracy on AR1, 97.4% on AR2, and 95.2% on all 9 features with threefold cross-validation scheme.

Chapter 3: Proposed Work

CNN model is used for classification process that can accurately classify a histology image as benign or malignant . Using Keras, we'll define a CNN (Convolutional Neural Network), call it Cancer Net, and train it on our images. The steps involved in the proposed method are described below, and it is represented in Fig. 2.

1. The IDC dataset is preprocessed to remove the noise from the instances.
2. We'll use classifier to train on 80% of a dataset and we'll keep 10% of the data for validation
3. Classify the dataset using deep neural network (CNN).
4. We'll then derive a confusion matrix to analyze the performance of the model.

Chapter 4: Methology

4.1 Preprocessing

The available dataset is random and there is no data label yet, so the initial stage of preprocessing is to make improvements by sorting and labeling data and explaining the types of processes that process raw data to prepare for other process procedures. In this dataset there is still incomplete data or missing value denoted by the "?". Thus, data refinement is required by using data cleaning technique to fill missing values on the dataset by handling using the average attribute values of all samples residing in the same class.

4.2 Feature Selection

The importance of feature selection in a machine learning model is inevitable. It turns the data to be free from ambiguity and reduces the complexity of the data. Also, it reduces the size of the data, so it is easy to train the model and reduces the training time. It avoids over fitting of data. Selecting the best feature subset from all the features increases the accuracy. Some feature selection methods are wrapper methods, filter methods, and embedded methods.

4.3 Classification

Splitting of the dataset is done randomly without following any sequences. After partitioning, a training set of data is initially applied to the classifier. This deep neural network classifier has an input layer with four input nodes, three hidden layers with 32,64,64,128,128 and 128 hidden nodes, and an output layer with a single node. Since this network has multiple layers with a huge amount of inner nodes, computationally it is expensive but provides promising results after training the model.

4.3.1 Convolution Neural Network

Technically, deep learning CNN models to train and test, each input image will

pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.

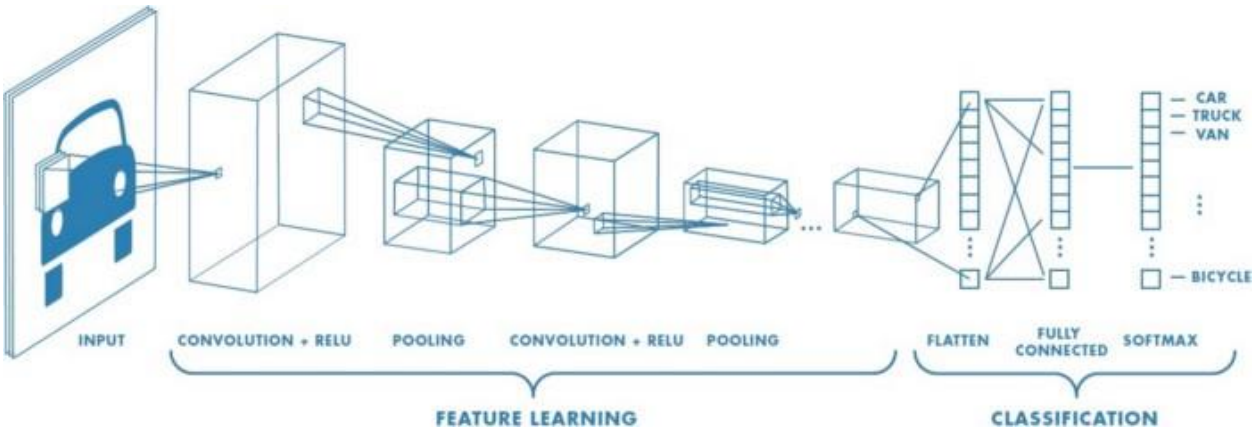


Figure 1: Neural network with many convolutional layers

Convolution Layer

Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel.

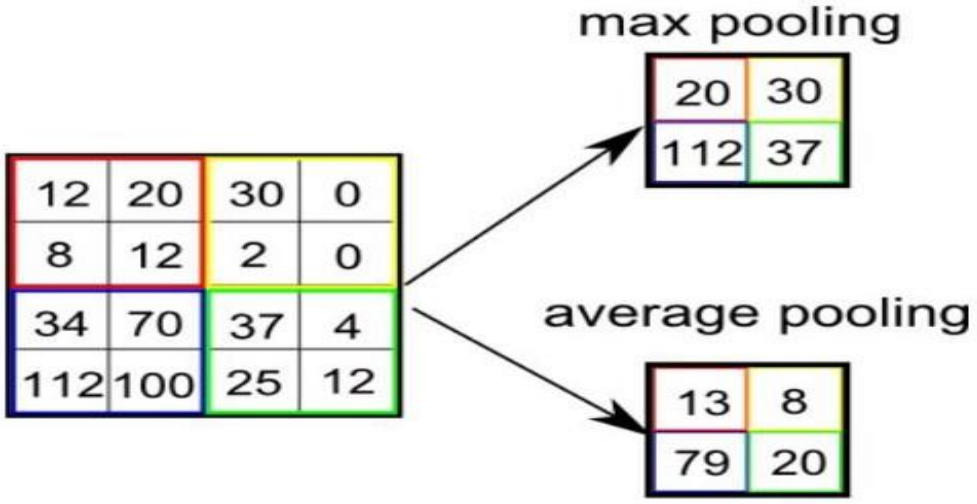
Padding

There are two types of results to the operation — one in which the convolved feature is reduced in dimensionality as compared to the input, and the other in which the dimensionality is either increased or remains the same. This is done by applying **Valid Padding**, or **Same Padding**.

Pooling Layer

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to **decrease the computational power required to process the data** through dimensionality reduction. Furthermore, it is useful for **extracting dominant features** which are rotational and positional invariant, thus maintaining the process of effectively training of the model.

There are two types of Pooling: Max Pooling and Average Pooling. **Max Pooling** returns the **maximum value** from the portion of the image covered by the Kernel. On the other hand, **Average Pooling** returns the **average of all the values** from the portion of the image covered by the Kernel.



Fully Connected Layer

The layer we call as FC layer, we flattened our matrix into vector and feed it into a fully connected layer like a neural network.

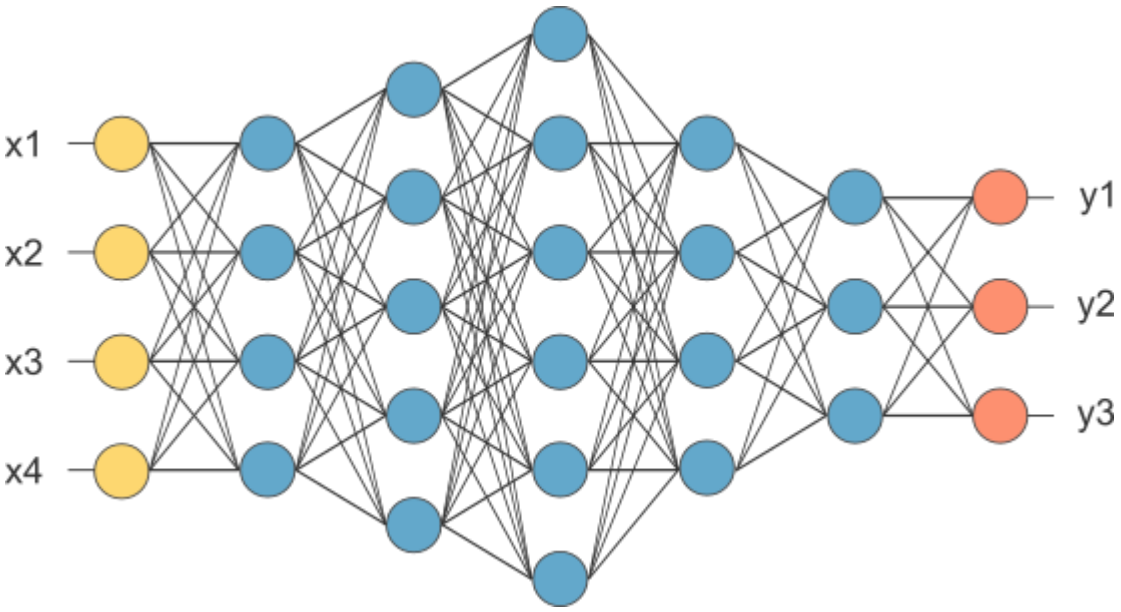


Figure 2: After pooling layer, flattened as FC layer

In the above diagram, the feature map matrix will be converted as vector (x1, x2, x3, …). With the fully connected layers, we combined these features together to create a model. Finally, we have an activation function such as SoftMax or sigmoid to classify the outputs as benign and malignant.

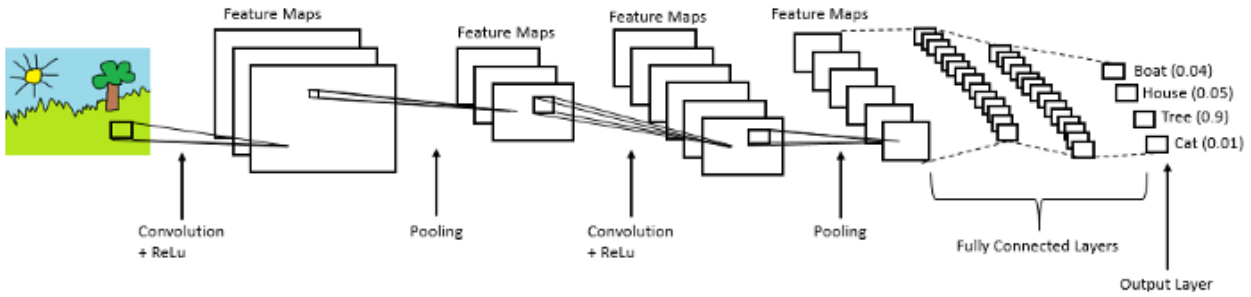


Figure 3: Complete CNN architecture

Chapter 5: Experimental Result

We are using the IDC_regular dataset (the breast cancer histology image dataset) from Kaggle. This dataset holds 2,77,524 patches of size 50×50 extracted from 162 whole mount slide images of breast cancer specimens scanned at 40x. Of these, 1,98,738 tests negative and 78,786 tests positive with IDC.

The performance of a model is estimated through confusion matrix. The confusion matrix helps to find the classified and misclassified rate of the system.
Effectiveness and performance of a system can be measured by calculating the accuracy.

CONFUSUSION MATRIX	ACTUAL	
PREDICTED	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

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

$$\text{F1-Score} = \frac{2*Precision*Recall}{Precision+Recall} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

Classification Report of proposed work

Classification Report:	Precision	Recall	f1-score	Support
0	0.91	0.87	0.89	39736
1	0.72	0.79	0.75	15769
avg / total	0.86	0.85	0.85	55505

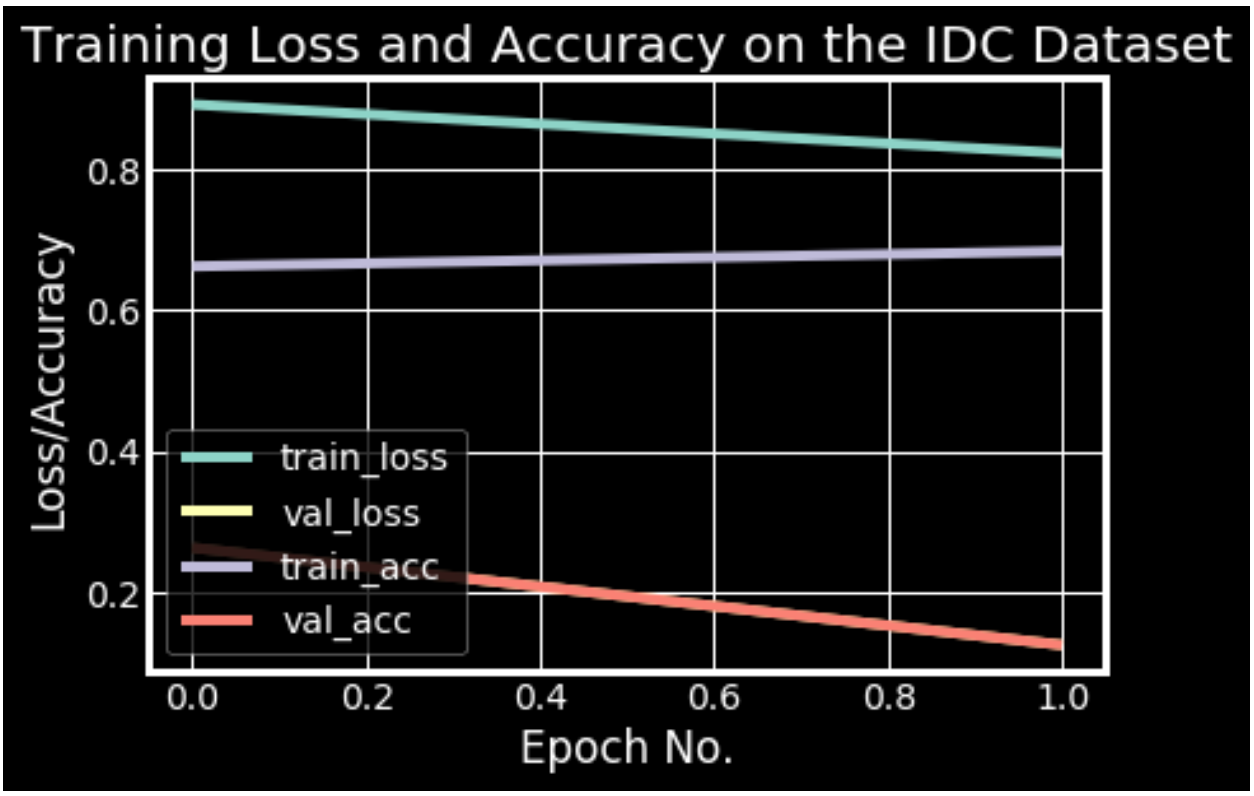
Confusion Matrix of proposed work

Confusion Matrix:	Predicted Benign	Predicted Malignant
True Benign	34757	4979
True Malignant	3271	12498

Three cases, although malignant, are predicted as benign

Output:

- Accuracy Score: **0.85**
- Specificity: **0.79**
- Sensitivity: **0.87**


Training Loss and Accuracy on the IDC Dataset
Loss/Accuracy
0.8
0.6
0.4
0.2
train_loss
val_loss
train_acc
val_acc
0.0
0.2
0.4
0.6
0.8
1.0
Epoch No.

Chapter 6: Conclusion

In this modern era, lots of people are facing many problems with modern age diseases. Breast cancer is one of the most common types of deadliest disease raising over time among different countries. Lack of awareness and post-identification of disease will be the major reason for more death rates. Computer-aided diagnosis will be a perfect solution for all kind of peoples to diagnose with accurate results. CAD system will not be a perfect replacement for professional doctors, but this aid will help them a lot, by assisting practitioners, to make a perfect decision by analyzing patient reports. Sometimes, practitioners may do some mistake due to lack of experience or poor analysis of reports. So, it will act as a better remedy for the current medical environment. More accurate decisions are taken, only if the model used to train the system will be unambiguous.

This system provides better results, compared with previous models, and needs a little improvement. The limitation of this system is training time of the algorithm since it has deeply trained the neural network.

Chapter 7: Future Work

For future research, we will create an automatic diagnosis and classification model for other disorders Furthermore, we will implement cloud computing that can be applied in the medical field to make it useful for the medical community and can expand this idea to Android-based mobile devices that can help patients to get their diagnostic results quickly

Chapter 8: References

[1] R. F. Malik, D. Z. Abidin, A. Zarkasi, Y. N. Kunang, J. S. Nurmaini and F. , "Breast Cancer Classification Using Deep Learning," *Research Gate,* pp. 237-240, 2018.

[2] K. Sekaran, "Breast Cancer Classification Using Deep Neural Networks," *Research Gate,* pp. 228-240, 2018.

[3] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder and R. M. &. W. , "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," 2019.

[4] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks--the ELI5 way," 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[5] Prabhu, "Understanding of Convolutional Neural Network(CNN)--Deep Learning," [Online]. Available: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148.