A project report on

An Integrated Detection and Treatment Recommendation Framework for Breast Cancer using Convolutional Neural Network and TOPSIS

Submitted in partial fulfillment for the degree of
Bachelor of Technology
in
Computer Science and Technology

Submitted by
Devdatta Basu (1612006)
Sheetal Kashid (1612020)

Under the guidance of
Dr Sanjay Pawar
Dr Debabrata Datta (External Guide)

Department of Computer Science and Technology
Usha Mittal Institute of Technology
SNDT Women's University

2019-2020

Approval Sheet

This is to certify that Devdatta Basu and Sheetal Kashid have completed the project report on the topic "An integrated detection and treatment recommendation framework for breast cancer using convolutional neural networks and TOPSIS" satisfactorily in partial fulfillment for the Bachelor's Degree in Computer Science and Technology under the guidance of Dr Sanjay Pawar during the year 2019-2020 as prescribed by Usha Mittal Institute of Technology, SNDT Women's University.

Guide

Dr Sanjay Pawar

Head Of Department

Dr Sanjay Pawar

Principal
Dr Sanjay Pawar

Examiner 1

Examiner 2

Acknowledgement

We take this opportunity to express our gratitude to all those who have contributed to our project. We extend our gratitude to our project guide, principal and head of department, Computer Science and Technology, Dr Sanjay Pawar for supervising us to make our project more credible, always attending to our concerns and evaluating our progress at every stage. We thank our mentor and guide Dr Debabrata Datta for his guidance and expertise in the field of Data Science which served as the basis for the recommendation module. Our heartiest gratitude also goes out to Dr Asawari Lautre, radiologist at Tata Memorial Centre, Mumbai for her suggestions and medical insight on the project, without which our project altogether would have been incomplete. We thank Dr Sonali Jadhav for her advice and for making sure that we got the best resources at Tata Memorial Centre. We also thank our project coordinator Prof Sonali Bodekar for arranging all the resources so that the whole process goes smooth and for addressing and being the conveyor of all important updates. We thank Dr Sudhir Bagde for his motivation and input during most of our phase examinations. We are thankful to all the teachers as they have been an integral part of our learning process. We thank our non-teaching staff for being there. We also acknowledge our friends and family for their constant love and support.

...

Date:

Devdatta and Sheetal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/ data/ fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Devdatta Basu
1612006

Sheetal Kashid
1612020

Date

Abstract

Breast cancer is the most prevalent type of cancer, accounting for 14% of all cancers among women in India, according to National Health Portal. Various machine learning including deep learning methods find extensive applications in the field of medicine i.e. Computer Aided Diagnosis (CAD) and one of its major examples is detection of cancerous cells. While many such systems explore the idea of classifying the region of abnormality as benign or malign, traditional treatment recommendation solely depends on the knowledge of the physician examining. This paper explores regression using Convolutional Neural Networks (CNN) to pin-point the probable abnormal region in mammograms and hence calculate several morphological, texture and histogram features associated with it. These features, in turn, are leveraged to recommend the next probable treatments. The recommendation is based on the extracted features and therefore constitutes a Multi-Criteria Decision Making (MCDM) problem. To implement this type of decision making rightfully, the paper makes use of the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) where the obtained features are used as criteria and expert opinions are used as alternatives to obtain ranked recommendations. The methodology proposed in this paper is capable of recommending the correct treatment with an accuracy of 81.5%. The proposed methodology would make treatment recommendation reachable even at the most remote places where advanced facilities including consultation with specialists are not easily available.
Keywords: *Tumor detection, Computer-aided diagnosis, Region of interest detection, Deep learning, MCDM, TOPSIS*

Contents

List of Figures

Nomenclature

TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
CNN	Convolutional Neural Network
MCDM	Multi-Criteria Decision Making
ANN	Artificial Neural Network
r	Roundness
p	Perimeter
A	Area
a	Acreage Ratio
MEC	Minimum Enclosing Circle
I	Inverse Moment
(i, j)	Indices
$P(i, j)$	(i, j)th Pixel Intensity
E	Energy
S	Entropy
C	Contrast Coefficient
μ	Mean
N	Image Resolution
σ^2	Variance
X_{sc}	Min-Max Scaled Data Point

X	Data Point
X_{min}	Minimum Data Point Value
X_{max}	Maximum Data Point Value
X_{inorm}	Normalised Weight
X_i	Original Weight
d	Euclidean Distance
x_i	Point in Space
y_i	Point in Space

Chapter 1

Introduction

Breast cancer has become a major disease among women of all age across the world. Breast cancer is a condition where lumps form within the breast due to uncontrolled growth of cells. Breast cancer is treatable when detected early. If not, it can be a fatal disease.

For detection at early age, it is advisable to get several tests done on regular intervals. Mammography is a radiographic technique which uses low-energy X-ray (i.e. few KeV) to detect characteristic masses or microcalcifications in the breast tissue [21]. Mammography is usually preferred over other methods because it is relatively less expensive and also sensitive to small lesions. Any region, that is likely to be cancerous, is represented by an irregular and dense accumulation of mass.

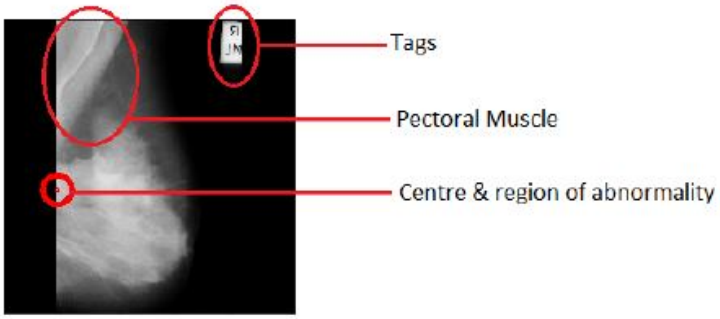


Figure 1.1: Components of a Mammogram

The main emerging technology that bolsters the process of interpreting medical images is Computer Aided Detection (CAD). CAD has the potential to help radiologists avoid overlooking a cancerous tumor or even decide what next course of treatment could be prescribed when reading a diagnostic mammogram. In particular, CAD algorithms are capable to detect the location of potential cancers and even the likelihood that a tumor is indeed cancerous. CAD systems have been proven to increase radiologists' ability to detect by approximately 10%. CAD algorithms continue to be a hot topic of research since it concerns real-time treatments, any failure to which may have preposterous outcomes [14].

A classical CAD algorithm to detect breast cancer comprises of mainly three steps as given below –

1. Preprocessing the images
2. Extracting the region of interest(s) (ROI)
3. Classification of tumor on the basis of features extracted – as benign or malign

Classic approaches to tackle tumor detection problems make use of several machine learning algorithms, such as [3] curvelet and Probabilistic Neural Network (PNN) classifier, [2] Naïve Bayes classifier and k-Nearest Neighbor (KNN) classifier and [4] compares the performance of Support Vector Machine (SVM) and Artificial Neural Network (ANN).

1.1 Motivation

Traditional breast cancer detection methods depend solely on the knowledge of the physician examining. However with the advent of new machine learning methods, nowadays it is possible to automatically classify if the tumor is benign or malign. This has given a source of relief even for the doctors as they can now get a second opinion, hence relieving them of the burden of detecting and classifying at sole discretion. However there are certain limitations to those models which only classify. Most of the times, even when the lump is benign, doctors choose to remove the lump altogether, just in case. For instance, if a patient uses the software and

finds out that the tumor she has is benign, only a doctor can say if it is likely to turn cancerous in the future. Here we find the need of a system that can say what the next step of treatment could be.

1.2 Intention

The proposed system takes existing CAD systems a step ahead and identifies the region of tumor as well. This region is used to extract morphological, histogram and texture features which hold key importance in determining the next course of treatment. After having found the features, the problem is to recommend the next probable course of treatment. This becomes a classic problem of multi-criteria decision making. Hence, we use the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [9] - the method which is based on Euclidean distance. TOPSIS has been successfully implemented on various systems like personnel selection and career matching where criteria played an important role in deciding the ideal solution.

1.3 Statistics and Analysis

The following data are obtained from International Agency for Research on Cancer, World Health Organisation [16]. In 2011, over 500,000 women succumbed to breast cancer. 50% of the cases and 58% of the deaths occur in less developed or developing countries. The survival chances in countries differs with different incomes. For instance, in countries like North America and Japan, the survival chance is 80%, in mid-income countries, the survival chance is 60% and finally low-income countries, the survival chance is 40%.

The probable reasons why the scenario in developing and under-developed countries is like what it is can be attributed to two main factors. First, lack of early-detection i.e. patients presenting at later-stages in the disease. Second, lack of adequate diagnosis and treatment facilities.

Cancer in breast outnumbers all other types of cancer by a huge margin worldwide (figure 1.2). In 2018, breast cancer accounted to a whopping 23.6% cancer-

related deaths among women i.e. almost a fourth (figure 1.3). As per projections made by International Agency for Research on Cancer's tools, India is likely to see a 61.2% hike in the total number of incident cases by 2040 (figure 1.4). This number, when compared with a low income country such as Africa (96.95%) (figure 1.5), although may seem promising, but comparison with the United States of America (25.8%) (figure 1.6) suggests otherwise. One of the reasons explaining the shallow hike in the number of cases in the USA can be attributed to the fact that the country currently holds the monopoly of cancer specialists and advanced equipment. The supporting figures are given below.

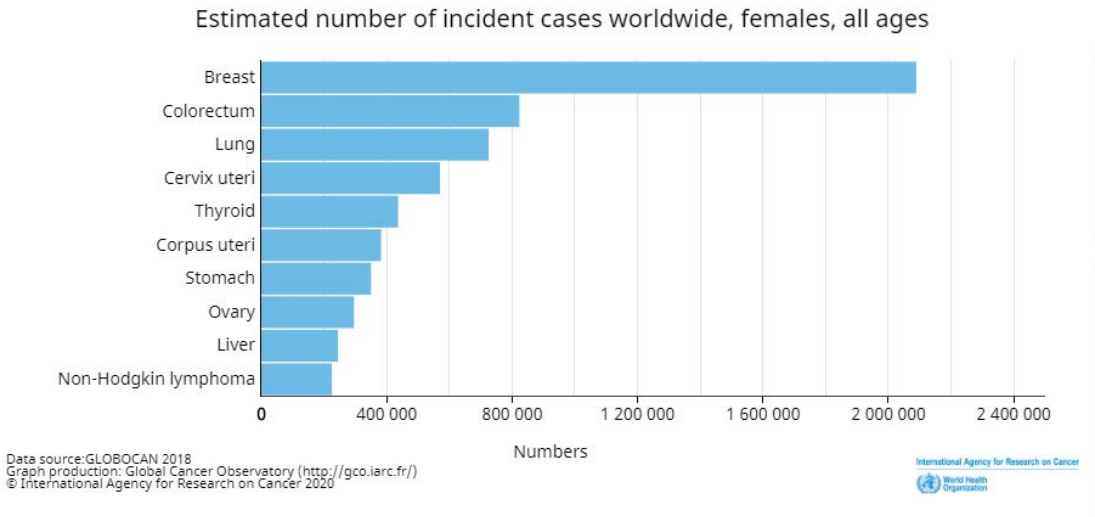


Figure 1.2: Estimated number of incident cases worldwide, females, all ages

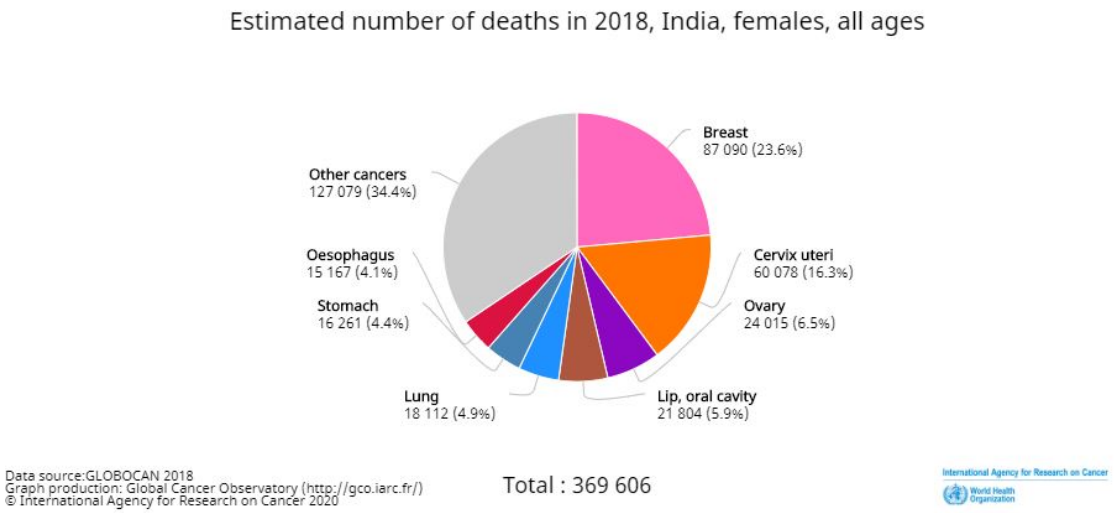


Figure 1.3: Estimated number of deaths in 2018, India, females, all ages

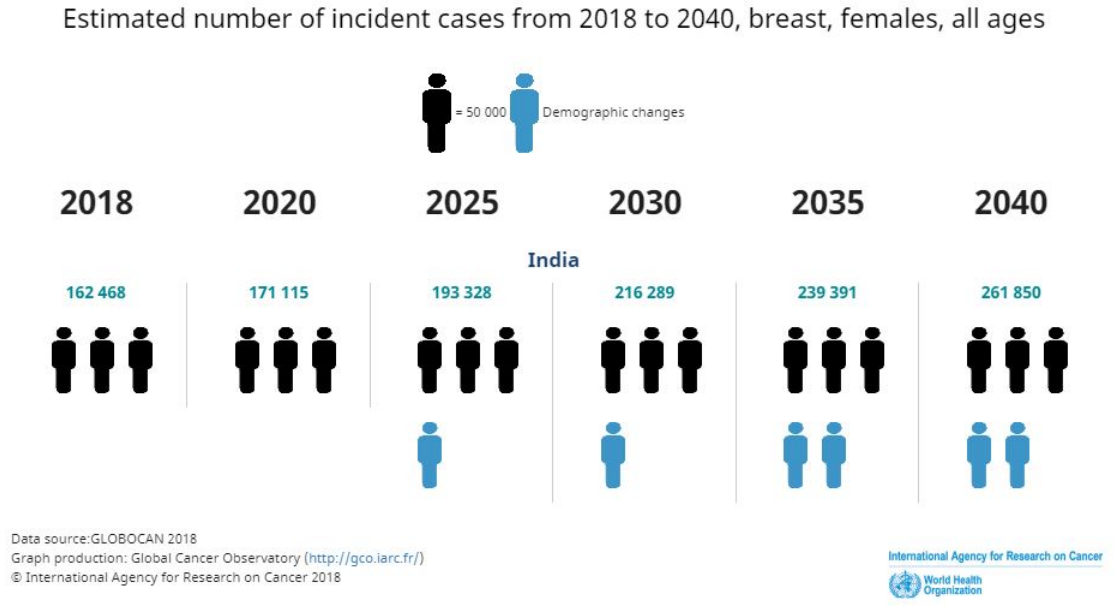


Figure 1.4: Estimated number of incident cases from 2018 to 2040, breast, females, all ages, India

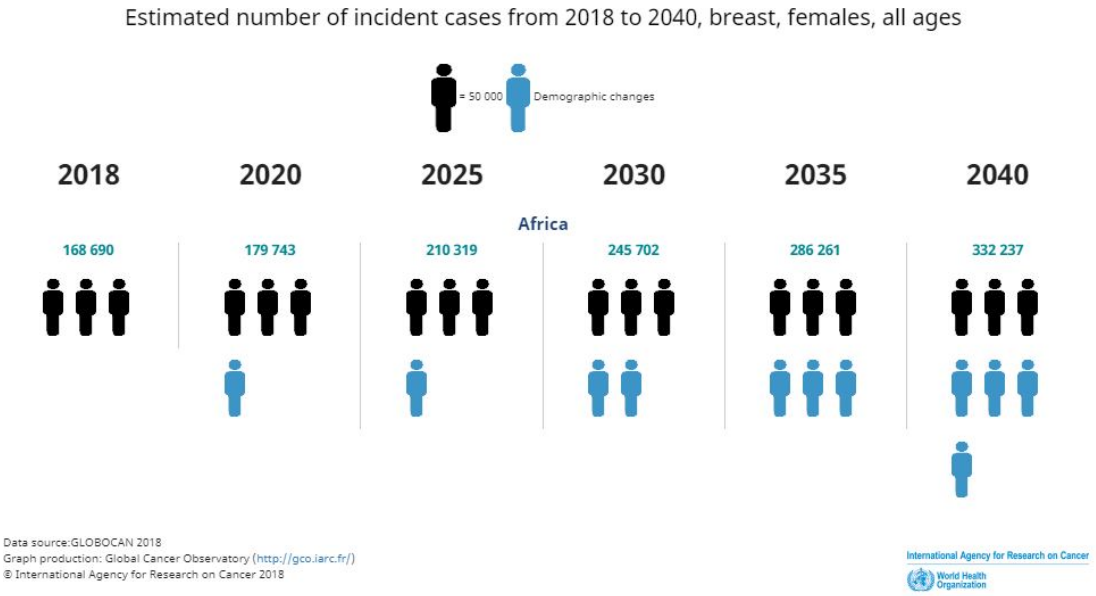


Figure 1.5: Estimated number of incident cases from 2018 to 2040, breast, females, all ages, Africa

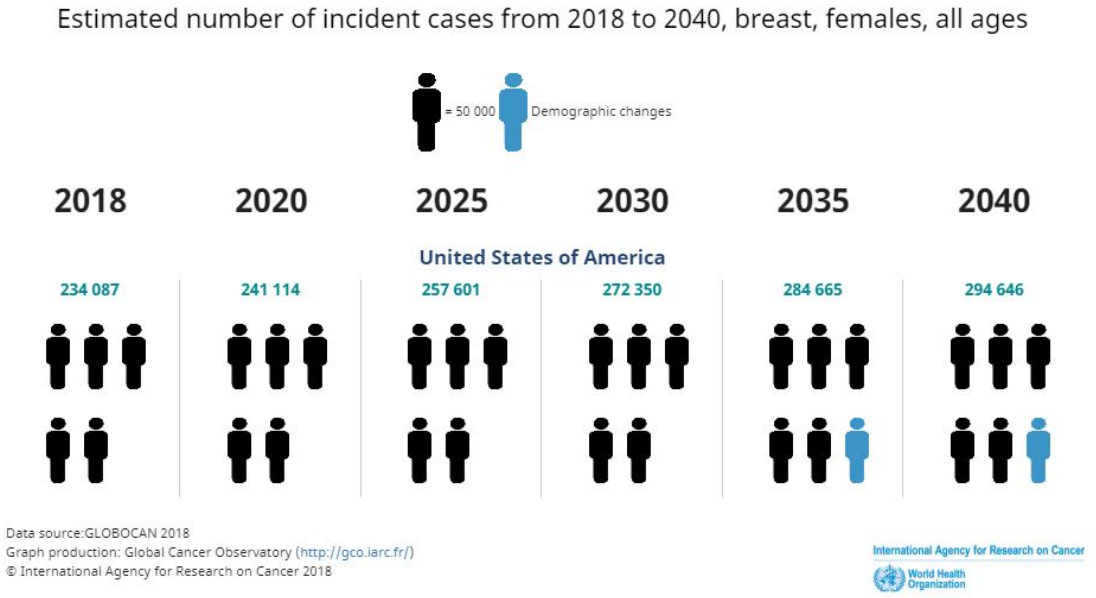


Figure 1.6: Estimated number of incident cases from 2018 to 2040, breast, females, all ages, United States of America

Chapter 2

Literature Review

Taheri [19] worked on a similar research on breast cancer classification with automatic thresholding utilising SVM and harris corner detection method. SVM was used to classify data, the algorithm generated a hyperplane which separated different classes and assigned every input of test dataset to one of the defined classes and automatic thresholding made the process independent from type of input images. [19] extracted features like intensity value, auto-correlation matrix value of detected corners and energy. The main idea was to make use of the difference in the intensity values of healthy and unhealthy part of the body and that they are overlaid. This method works well with precision of 96.8 and recall rate 92.5.

This research used the method of CNN feature extraction similar to a method Wang et al. [20] and Hu et al. [8] proposed for breast cancer detection, where deep features were calculated and Extreme Learning Machine (ELM) clustering was used for mass detection. A feature set containing deep, morphological, texture and histogram features was constructed. The ELM classifier was built with this set to classify the masses into benign and malign.

Lu et al. [12] proposed a method which utilized the median filter, contrast-limited adaptive histogram equalization and data augmentation method to preprocess the mammograms and train a classifier by using the CNN. In the preprocessing stage, to enhance the contrast of the images – CLAHE was used. The images of malignant tumors were flipped horizontally and vertically to generate more images with tumors present in different orientations. Median filter improved

the image quality and removed salt and pepper noise. They employed transfer learning method with Adam optimizer for fine tuning the CNN model. One-hot encoding was used to encode as malign or benign. Their model accuracy was 0.823 in testing dataset.

For the task of processing pectoral muscle, as an important ROI extraction step Rahimeto et al. [17] used Connected Component Labelling (CCL) method to make the process automatic. Tags were removed and pectoral muscle processing was automated using Otsu's multi-thresholding technique, CCL and the biggest blob extraction method. Pectoral muscle processing technique utilized one-third of its perimeter on the edges of the mammogram in order to extract a binary region.

M. A. Al-masni et al. [1] proposed You Only Look Once (YOLO) based CAD system consisting of four main stages - preprocessing the mammograms, feature extraction using multi-convolutional deep layers, mass detection with confidence model and finally classification using fully connected neural network. To train YOLO, the information of ROI masses and their types with a training set of mammograms was used. This trained YOLO-based CAD system detected the masses and classified them as benign or malign. This system detected the mass location with an overall accuracy of 96.33% and the benign malign lesions were distinguished with an overall accuracy of 85.52%. This system was capable of concurrently detecting and classifying the masses and also dealt with difficult cases of breast cancer, for example the mass lying in the dense regions such as pectoral muscle.

TOPSIS and Analytic Hierarchy Process (AHP) methods for decision-making were used by Lokare et al. [11] in order to select best course after Higher Secondary Certificate (HSC). The problem was seen as a Multi-Criteria Decision-Making (MCDM) problem i.e. there were total four criteria - interest, employment opportunity, duration and fee along with possible alternatives to choose the best career option. AHP was used to calculate the weight of each criterion. To measure the rank, both TOPSIS and AHP were employed. Results were compared and final ranks were assigned. It was found that the concept of TOPSIS is that the most preferred alternative should not only have the shortest distance from the positive ideal solution, but also have the longest distance from the negative ideal solution. AHP method was also used further to calculate the final ranking.

In the work done by Nabeeh et al. [13] they developed an application for personnel selection using Integrated Neutrosophic-TOPSIS. They divided the system into three stages. The first stage was Hierarchical Process Method which dealt with determining the objectives, criteria and alternatives and computing global weights of criteria which were considered to ensure that the candidate applicant fulfills the enterprise needs. The second stage defines the neutrosophic scales for criteria and alternatives for which they applied score function for changing neutrosophic scales into crisp values. In the third Stage TOPSIS was utilized and was applied to choose the ideal candidates by establishing positive and negative areas of candidates. After normalizing the judgements, Euclidean distance of the ideal solution was measured to compute relative closeness to praise alternatives based on which the ideal solution was chosen.

[5] introduces a study to build a comprehensive factor model for better career matching that integrates the typical and atypical factors. The model proposed identifies the key factors for career choice by taking advantage of TOPSIS and Fuzzy Cognitive Map (FCM), also considering the relationship between typical factors, atypical factors and occupational items. The results indicate a valuable research field towards career counseling and human resource planning and also can benefit the design and implementation of an operational career prediction system in the future.

[22] presents Convolution Neural Network (CNN) based face recognition method. Since it deals with location detection, the CNN model used in this paper could be used for the proposed system. This network consists of three convolution, two pooling, two full-connected and one Softmax regression layer. Stochastic Gradient Descent (SGD) algorithm is used to train the feature extractor and the classifier, which can extract the facial features and classify them automatically. The dropout method is used to solve the over-fitting problem. The Convolution Architecture For Feature Extraction framework (Caffe) is used during training and testing. The face recognition rate of the ORL face database and AR face database based on this network is 99.82% and 99.78%.

[10] compares two model networks of deep learning technique. The overall process involves image preprocessing followed by classification and performance evaluation. After evaluating the performance of deep learning models - VGG16

[18] and ResNet50 [7] which were used to classify between normal and abnormal tumor using IRMA dataset [15], the result showed that VGG16 produces the better result with 94% compared to ResNet50 with 91.7% in terms of accuracy.

Chapter 3

Architectural Overview

The product based on the research can be visualised as an application which takes mammograms and patient details (including medical test reports) as input from the user. Later, it recommends the next probable treatment to the patient which can be later discussed with a doctor just in case. In the backend, the input is preprocessed and the pectoral muscle is processed. Next, the system detects the coordinates and radius of the abnormality which forms the basis for feature extraction. The features are then appended to the database and are normalised. Using these features, ranked treatment recommendations are offered using the TOPSIS recommendation engine.

3.1 Technology Used

- Operating system - **Windows 10**
- Hardware - **NVIDIA GeForce MX150, NVIDIA GeForce 940MX**
- Environment - **Anaconda Jupyter Notebook, Google Colab**
- Preprocessing - **OpenCV Python library**
- Pectoral muscle - **OpenCV Python library**
- Locating the region of abnormality (CNN) - **Keras using Tensorflow v2.x backend**

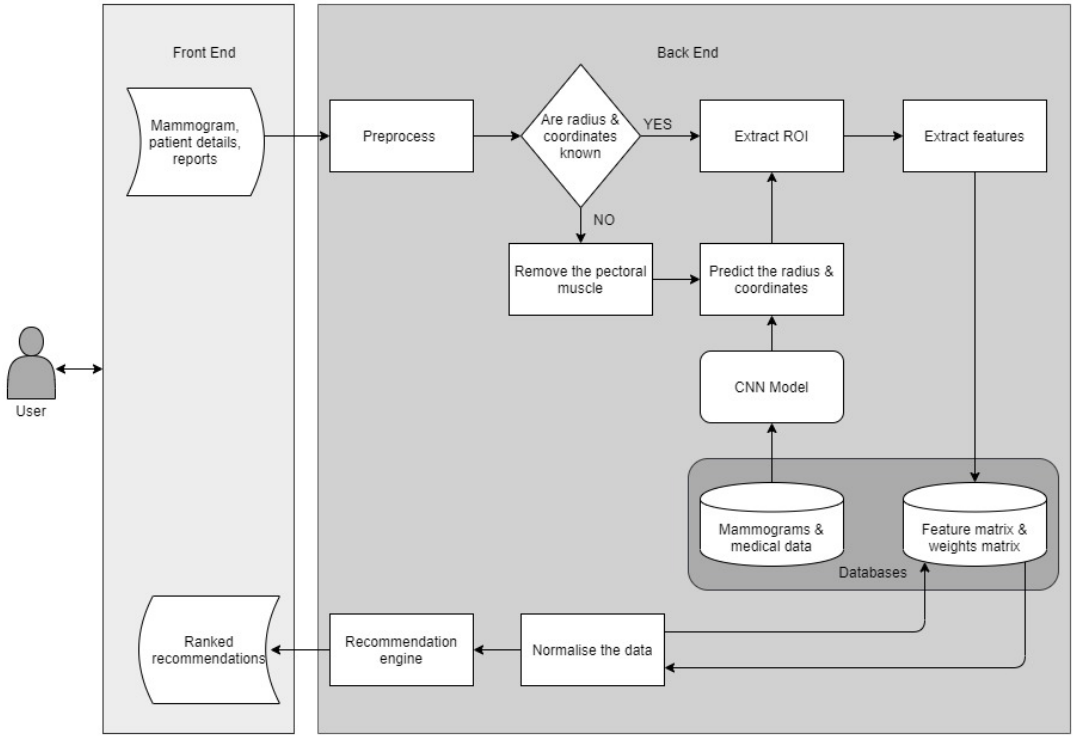


Figure 3.1: Architectural Overview

- ROI extraction - **OpenCV Python library**
- Feature extraction - **Pandas Python library**
- TOPSIS - **Numpy Python library**

Chapter 4

Design of the Proposed System

Healthcare, as we know it, is coherently dependent on professional medicos. Whether a patient can get his/ her appropriate treatment primarily depends on the availability of such medical professionals. In countries where the number of patients exceeds the number of professionals by a huge unthinkable margin, it can be inferred that many severe cases indeed go unattended. Here, the disease dealt with is breast cancer. Breast cancer has become a major disease among women of all age across the world and as research has proven, if not detected early, it can be fatal.

Though there exist several screening methods like thermal imaging, biopsy, histopathology, UGG, MRI, etc., the most preliminary imaging method is mammography. It is widely available in many pathological labs, hence making it more accessible as compared to other methods of imaging. Therefore due to its wide availability, mammogram is the ideal candidate. The main aim of the proposed system is to exploit deep learning methods and mathematics for the mammograms to rightly recommend the next course of treatment for the patient i.e., whether she should go for other types of imaging or re-imaging, immediately go for treatment or not necessarily go for any treatment, hence leaving room for integration of other screening methods and alternatives as well. Making a system like this reachable to the most remote places, in today's world, can help in early detection and can make treatment easier.

4.1 Shortcomings of Existing Machine Learning Models

Classic approaches to tackle tumor detection problems make use of several machine learning algorithms, such as [3] curvelet and PNN classifier, [2] Naïve Bayes classifier and KNN classifier and [4] compares the performance of support vector machine (SVM) and artificial neural network (ANN). Although machine learning methods are hugely successful in identifying patterns and trends in underlying data with high dimensionality and density, it struggles when applied to critical applications such as medicine. The cause for this can be attributed to the error susceptibility of machine learning algorithms, and how also these tend to work better when the volume of data is more. Another one of the shortcomings of learning algorithms is these fail to extract features from the given data.

While other existing systems like that of Wang et al. [20] classifies the tumor cells as benign or malign using CNN for feature extraction and Extreme Learning Machine for classification. The proposed system employs a similar structure which is using CNN for feature extraction but uses these features as a MCDM problem to find the probable methods of treatment. Another system classifies the tumor cells as malign or benign using YOLO based CAD system but the system does not provide any room for detecting different kinds of tumor. Taheri [19] worked on a similar research on breast cancer classification with automatic thresholding utilizing SVM and Harris corner detection. [19] extracted features like intensity value, auto-correlation matrix value of detected corners and energy. The main idea was to make use of the difference in the intensity values of healthy and unhealthy part of the body and that they are overlaid. The proposed system uses CNN for detecting the tumor considering various layers hence extracting more features. The number of features extracted by the research proposed by Taheri would have been insufficient for a good result while implementing and taking MCDM approach.

4.2 Design

When it comes to images, machine learning algorithms become even more weak. There are many reasons why it is so. First, a single image is in itself a huge amount of data. For instance, the dataset used in the proposed system is a set of 322 images each of resolution 1024*1024. Therefore, any system that is built around this data is supposed to process 337,641,472 data points. Second, since images are merely a 2D representation of the 3D world around us, there is a huge scope of information loss. Third, images are susceptible to noise, in the form of lens flare, artefacts, checkerboard effect, etc. Fourth, how the images are interpreted by the computer differs from how it is done by humans. For instance, human brain has a deeper sense of depth and perspective as compared to the computer, which sees images as a mere matrix of numbers.

Hence the approach usually taken while dealing with image processing problems is the use of deep learning, in particular, Convolutional Neural Networks (CNNs). A CNN typically consists of a combination of one or more convolution, pooling, fully connected and activation layers. The main essence of CNN is the convolution operation. A kernel slides over the entire image to detect aspects and assigns a heavy weight for regions with significant features. This weighted matrix (tensor) forms the input to the next layer and so on. This deep learning approach is often preferred over other machine learning techniques because of its ability to detect features even if the image data is skewed, zoomed, sheared, tilted, flipped, or transformed visually in any other way. This becomes a major help as tumors may widely vary in size and this property might also denote its severity.

The research intends to build an integrated framework consisting of detection and treatment recommendation modules. The approach does not directly base the research only on the visual aspects of the tumor images but also on various other features extracted from the region of interest. These include –

1. Morphological features such as roundness, roughness, etc.,
2. Texture features such as energy, entropy, etc., and
3. Histogram features such as mean, variance, peak, skew, etc.

This is achieved by training a convolutional net (CNN) to locate the centre of tumor mass (minimum bounding circle) and its radius. In order to locate the centre of tumor, regression instead of classification is used while running the CNN. Regression using CNN has found its applications chiefly in the domain of facial recognition. In such problems, multiple key points are to detected in each picture, given the x-y coordinates of such points. To train the system MIAS database was used, obtained from Kaggle. The MIAS database consists of 322 breast mammogram images and a file consisting of the x-y coordinates of the suspicious region along with its radius. The image dataset is employed as the independent entity and x-y coordinates and radius as the predictor entity.

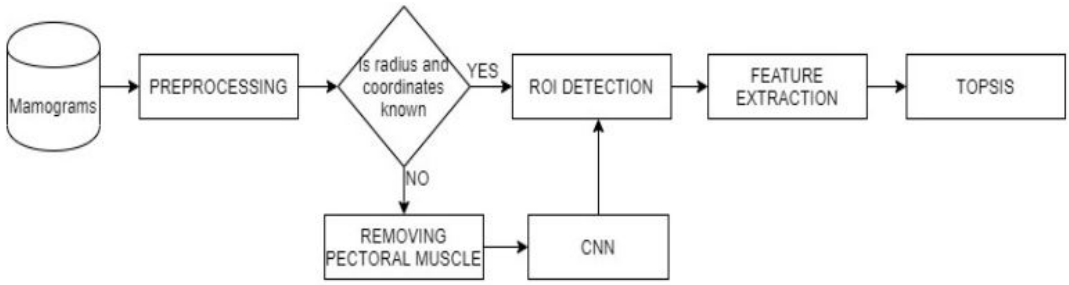


Figure 4.1: Block Diagram

All 17 relevant features are calculated using the centres and radii. These are potent enough to pose as the criteria for decision-making, hence turning this problem into a classic case of Multi-Criteria Decision-Making (MCDM). The best-known algorithm to deal with such problems is the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [9]. TOPSIS is preferred because of its simplicity, rationality, comprehensibility, good computational efficiency and ability to measure the relative performance for each alternative in a simple mathematical form. The primary intention of the system proposed is to recommend the next probable course of treatment for the patient i.e. whether the patient should –

- Go for screening again,
- Go for treatment immediately and
- Not go for treatment

4.3 Modules

4.3.1 Preprocessing

Input Mammograms

Output Preprocessed mammograms

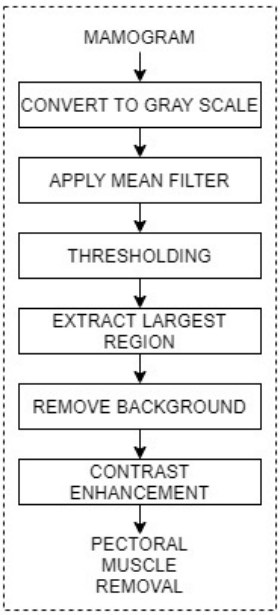


Figure 4.2: Preprocessing

4.3.2 Image Processing of Pectoral Muscle

Input Preprocessed images

Output Pectoral muscle processed mammograms

First the orientation is identified then the right-oriented images are flipped and the dark border region on the left is cropped out. This is done because the algorithm developed for removal makes an assumption that the pectoral region lies in the top left corner of the image.

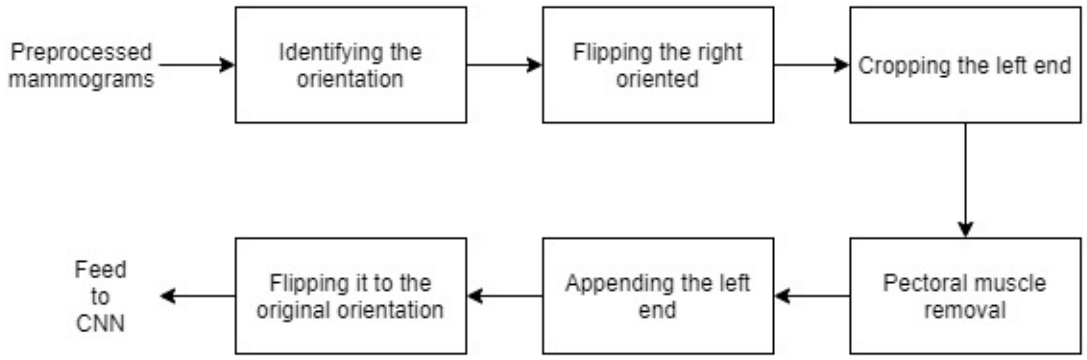


Figure 4.3: Image Processing of Pectoral Muscle

4.3.3 Detection

Input Pectoral muscle processed mammograms and description data

Output Radius and coordinates of abnormality

4.3.4 ROI Detection

Input Radius and coordinates of abnormality

Output Contour region

4.3.5 Feature Extraction

Input Contour region

Output Various features

4.3.6 Recommendation

Input Feature matrix of mammogram and weight matrix for corresponding feature weights

Output Ranked recommendation

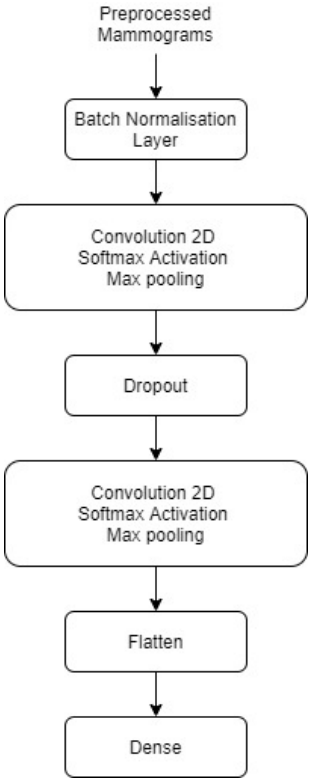


Figure 4.4: CNN

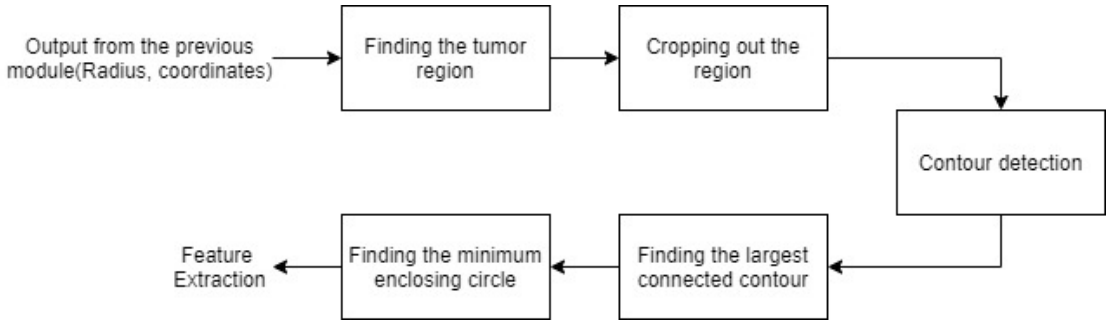


Figure 4.5: ROI Detection

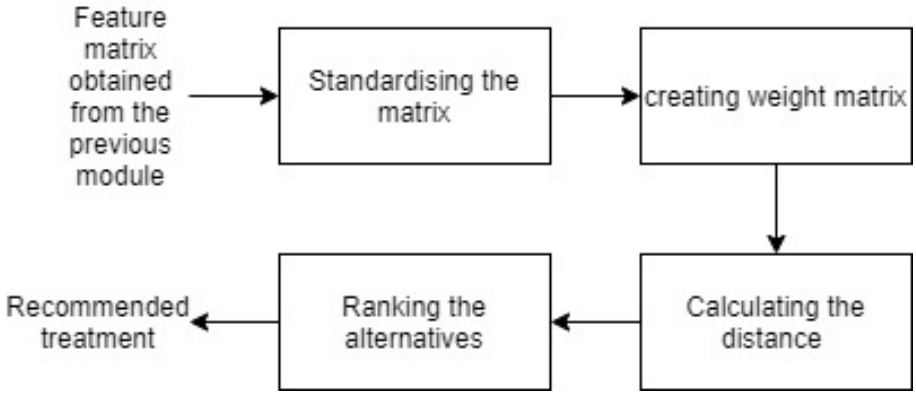


Figure 4.6: Recommendation - TOPSIS

Chapter 5

Implementation of the Proposed System

Like related papers, the system proposed also consists of seven main steps – image acquisition, mammogram preprocessing, image processing of pectoral muscle, detecting the tumor, Region of Interest (RoI) extraction, feature extraction and recommendation. In image acquisition, for training purpose appropriate labelled images are searched for. In image preprocessing, noise and irrelevant portions (like tags) are eliminated from the images and contrast of the images are enhanced, thus making the image suitable for further processing. Following image processing of pectoral muscle, the processed mammograms is convolved with apt kernels for locating the centre of the tumor and finding its radius. The tumor region is further refined using contours. Several features like morphological features, texture features and histogram features are extracted. These extracted features represent the criteria based on which a recommendation is to be made. Expert opinions form the alternatives that act as the basis of decision-making for this multi-criteria problem. Furthermore, TOPSIS is applied for detailed ranking of the probable recommendations that the patient can go for.

5.1 Image Acquisition

To carry out the project, MIAS mammography images from Kaggle were used. The dataset consists of 322 mammogram images of 1024 x 1024 size arranged in pair of films where each pair represents the left (even filename numbers) and right mammograms (odd filename numbers) of a single patient. The image matrix contains the mammogram at its centre. A coordinate system is defined around the bottom-left corner. The dataset also contains a txt file, which provides the labels for the images.

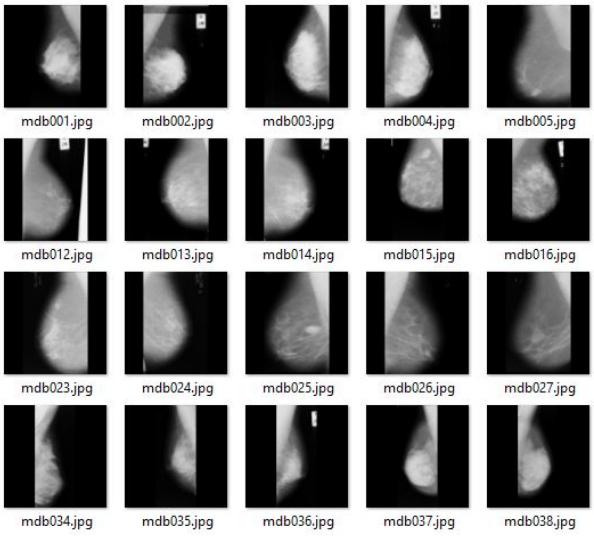


Figure 5.1: Images

REFNUM	BG	CLASS	SEVERITY	X	Y	RADIUS
0	G	CIRC	B	535.0	425.0	197.0
1	G	CIRC	B	522.0	280.0	69.0
2	D	NORM	NaN	NaN	NaN	NaN
3	D	NORM	NaN	NaN	NaN	NaN
4	F	CIRC	B	477.0	133.0	30.0

Figure 5.2: Text file

5.2 Image Preprocessing

The raw mammogram images deceptively are in RGB which basically sets the apparent grayscale images into three channels. To the human perception, however, these images are indeed in grayscale. The images also contain noise like tags or other borders due to improper scanning. This noise is generally in the form of high frequency regions. In the further steps when we apply CNN, these high frequency regions may add to the weights supplied to the neurons. The subjective quality of breast mass is often not visible for human interpretation. Contrast is defined as the difference in brightness of two adjacent surfaces. Oftentimes, the contrast difference helps in separating the breast tissues (foreground) from the rest of the breast (background).

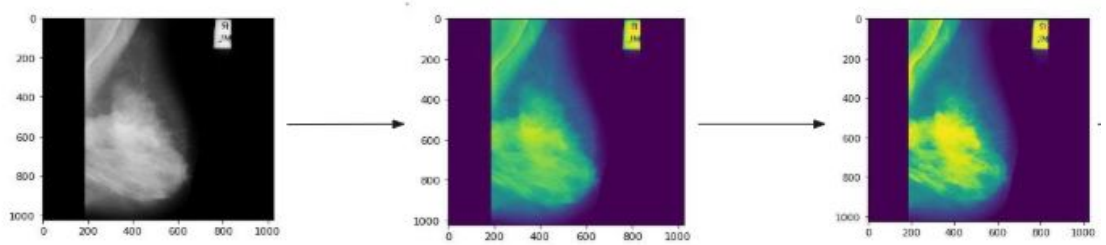


Figure 5.3: Preprocessing

To remove tags and other background disturbance present in the grayscale images, thresholding is done so as to separate the foreground from the background, obtaining a mask, as shown in figure 5.4. As evident, the breast region lies within the largest portion of the masked image.

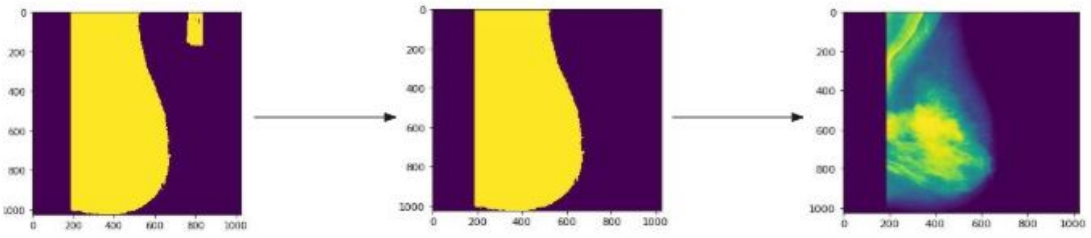


Figure 5.4: Removal of Tags

5.3 Image Processing of Pectoral Muscle

Pectoral muscle is what connects the front of the human chest with the bones of the upper arm and shoulder. It lies under the breast and hence is often captured in mammograms. Pectoral muscle forms a region of high frequency which might make unnecessary contribution to the working of CNN. Hence it becomes necessary to process such regions before proceeding. Following are a few points to be considered while dealing with pectoral muscle –

- It usually lies on the top left corner (in case of left-oriented breasts) or top right corner (in case of right-oriented breasts) of the mammogram.
- It is a region of very high and consistent frequency.
- It usually has a shape of an inverted right-angled triangle.
- The angle that the boundary of pectoral muscle makes with the boundary of the image is generally between 30 and 70 degrees.

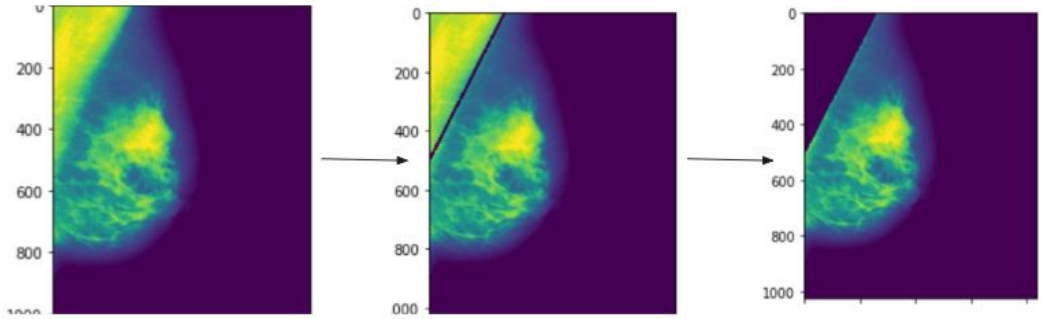


Figure 5.5: Linear Cut-Off Method

The inferences that can be drawn from the above observations are –

- For images -
  - In case of padded images, the padding must be removed in order to process (padding will cause the pectoral muscle difficult to locate).

  - For simplicity and reuse of code, all right-oriented breast images must be identified and flipped.

- The deviation of frequencies within adjacent pixel values must be very small.
- The pectoral muscle makes a straight line with angle lying somewhere in between 30 and 70 degrees.

Although there exists a method to process the pectoral muscle part, it is not very efficient as it considers the pectoral muscle to be exactly linear in shape. In real images it is hardly so. Therefore, having a linear cut-off may remove some of the breast mass as shown in figure 5.5. This method also poses a problem when trying to identify the start and end of the line that cuts it off. If the identification is not done properly, the linear method may fail.

The method we propose yields a better result as far as the coverage of breast mass is concerned because ours deals with pixel intensity of all the breast mass instead of just two pixels to identify the start and end of the line. For the algorithm to function, there is a need of all input being strictly left oriented. To do so an automatic orientation detection method was applied, which detected the orientation of the images with 100% accuracy for the database in discussion. After having flipped the right oriented breast mammograms, there is a need to crop out any additional columns of pixel that may be there due to the imaging device. This was done with an accuracy of 100% for the database in discussion. The proposed method is shown in algorithm 1.

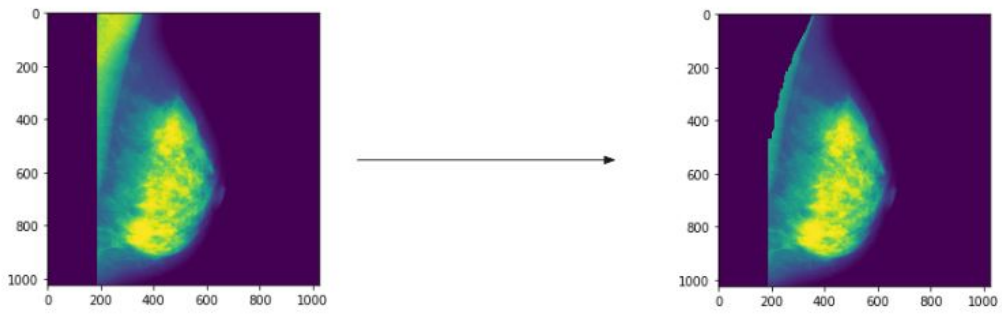


Figure 5.6: Proposed Method

Algorithm 1: Image processing of pectoral muscle

Input: Preprocessed images

Result: Images containing breast mass

Procedure;

for *each image* $\boldsymbol{i}$ **do**

Consider the first $\mathbf{N}$ rows from the top right corner, traverse each row and locate the index $\mathbf{x_j}$ of the last pixel having value greater than threshold $\mathbf{T}$ where;

1. $0 \leq \mathbf{T} < 255$
2. $1 \leq \mathbf{j} < \mathbf{N}$

Consider the first $\mathbf{N}$ columns from the top right corner, traverse each column and locate the index $\mathbf{y_j}$ of the last pixel having value greater than threshold $\mathbf{T}$ where;

1. $0 \leq \mathbf{T} < 255$
2. $1 \leq \mathbf{j} < \mathbf{N}$

Set $\mathbf{X}$ = maximum $(\mathbf{x_j})$ and $\mathbf{Y}$ = maximum $(\mathbf{y_j})$;

Set $\mathbf{Z} = \mathbf{X}$;

for *each row* $\boldsymbol{i}$ *in image such that* $\boldsymbol{i} < \boldsymbol{Y}$ **do**

Select the row and mark the last pixel with pixel value greater than $\mathbf{T}$ having index $\mathbf{k}$ such that $\mathbf{k} \leq \mathbf{Z}$;

Cut off the region along the row from indices 0 to $\mathbf{k}$;

Set $\mathbf{Z} = \mathbf{k}$;

end

end

5.4 Detection of Centre the Radius of Abnormality

As it is evident, to calculate most of the features, it is necessary to locate the centres and radii of the abnormalities which we do with the help of convolutional neural network. Finding numeric quantities from an image forms a classic example of regression. One of the most famous use cases of regression using CNN is keypoint detection. Here, given the coordinates of various such points, the challenge is to train the network in such a way that it is able to automatically identify the points. The problem here, concerning the mammograms and abnormalities is similar to the above-mentioned use case i.e. to train the net to automatically locate the point. The CNN model used is shown in figure 5.7.

5.4.1 Why CNN and not ANN?

A CNN is an algorithm which takes in an input image, assigns importance - weights and biases - to various features in the image and can differentiate one from the other or extract features. A CNN is successfully able to capture spatial and temporal dependencies in an image as a result the architecture fits better to image dataset. One of the other reasons why CNN is more suited for image datasets is with CNN, there is a sharp reduction in number of parameters involved (due to convolution operation) and also weights are reused.

5.4.2 Batch Normalisation

Training most deep neural networks can be challenging as the network can be sensitive to initial weights and configuration. One of the possible reasons for this is the distribution of inputs to layers deep in the network may alter for every batch as the weights are updated. This can cause the algorithm to chase a moving target.

Batch normalisation layer is used for training very deep neural network models. It standardises the inputs to the next layers for each mini-batch. This stabilises the learning process and reduces the total number of epochs required for training.

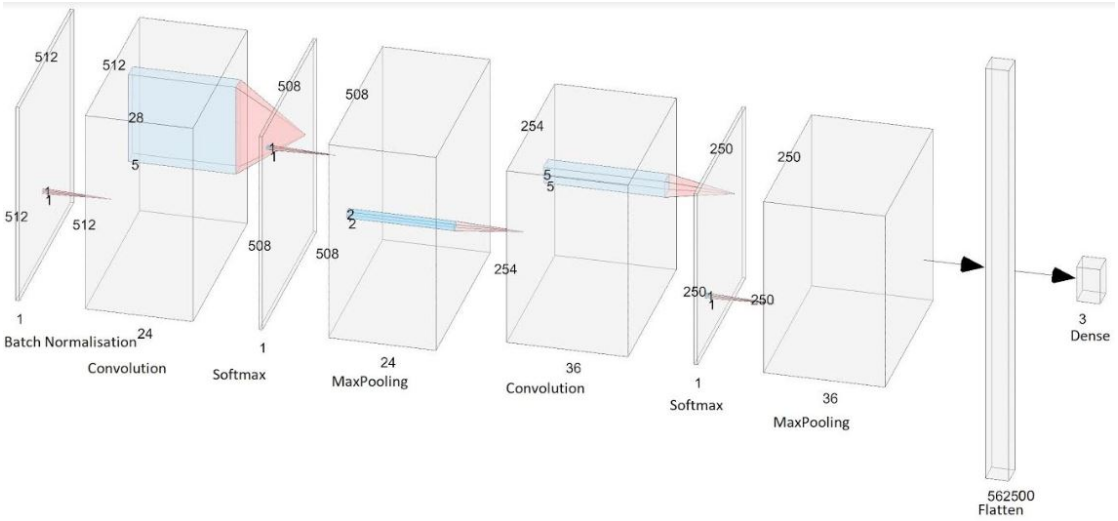


Figure 5.7: CNN Regression

5.4.3 Convolution

Convolution layer is responsible for feature extraction. It can be thought of as an ordered element-wise multiplication of two matrices to obtain a value. The first convolution layer looks for basic features such as straight lines, dots, diagonals, etc. The second convolution layer uses these features to look for more complex ones like curves, circles, polygons, etc. As the number of convolution layers increases, more complex features come into picture. What the convolution layers in this application tries to do is -

- Locate the centre
- Build consecutive outward-growing rings to find the extent to which the region goes i.e. radius

5.4.4 Max Pooling

Max pooling is a discretisation based process. The objective is to down-sample an input representation (image matrix), reducing its dimensionality and enabling inferences to be made about features contained in the conglomerated regions.

5.4.5 Activation - Softmax

The equation for Softmax is -

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

Softmax is mainly used to normalise network output to fit between 0 and 1. It is used to represent the certainty i.e. probability in the network output. It is useful because it converts the output of the last layer into what is essentially a probability distribution.

5.4.6 Flatten

Flatten is used to convert the n-dimensional matrix to a 1-dimensional array.

5.4.7 Dense

A dense layer is nothing but a linear operation on the input vector to find out the probability of the predicted output.

5.5 ROI Extraction

The region obtained using the centres and radii predicted by the CNN model is cropped and an approximate circular contour is obtained as shown in figure 5.8.

Inside this probable region, exists the actual tumor region/ abnormality. In our visit to a cancer institute, Dr Asawari Lautre, radiologist at Tata Memorial Centre, Mumbai marked the exact abnormal lump/ tumor (shown in figure 5.9 left), which in turn is kept as the benchmark when trying to automatically mark the contour around the abnormality.

The result obtained after setting the appropriate contours according to benchmark is shown in figure 5.9 centre. Thereafter, an additional minimum enclosing circle (MEC) is drawn around the abnormality detected as shown in figure 5.9 right.

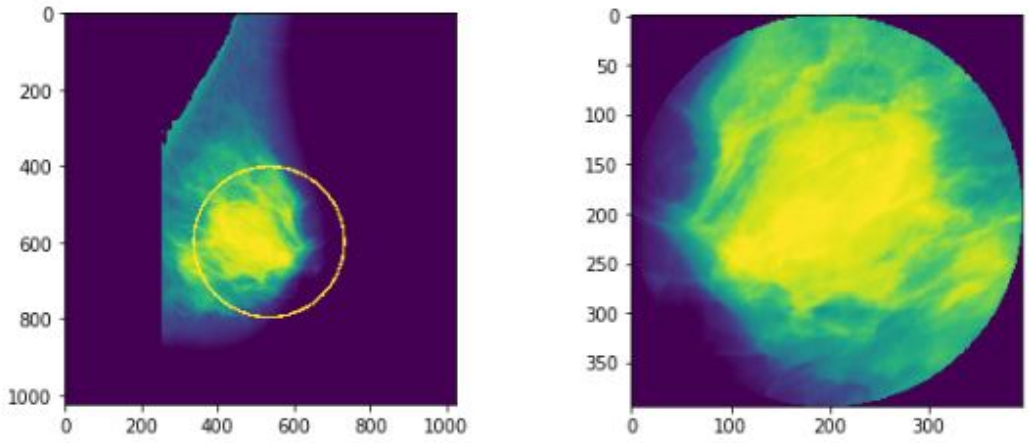


Figure 5.8: Cropping

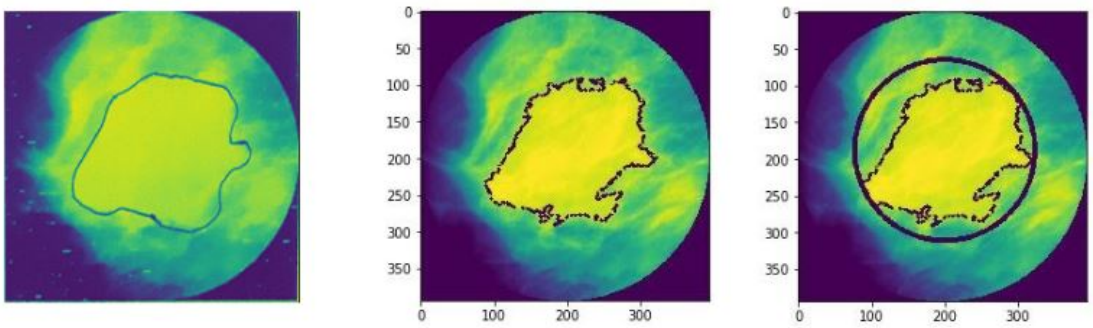


Figure 5.9: Abnormality Marked (L), Automatically Detected (C), Minimum Enclosing Circle (R)

5.6 Feature Extraction

The features used in the TOPSIS recommendation module can be broadly classified into -

- Morphological
- Texture
- Histogram

5.6.1 Morphological Features

The morphological features used along with their significance are listed below -

Roundness

Roundness signifies how much similar the shape of the tumor is to an ideal circle. The metric typically lies between 0 and 1 (appendix: A.1).

Acreage Ratio

Acreage ratio is also known as land-to-building ratio. In our context it signifies how much area of an ideal circle does the tumor occupy (appendix A.2).

5.6.2 Texture Features

The texture features used along with their significance are listed below -

Energy

Energy signifies how much the neighbouring pixels vary from each other (appendix A.4).

Entropy

Entropy signifies how irregular the texture of the tumor is (Appendix A.5).

Contrast Coefficient

Contrast coefficient signifies how bright is the tumor from its surrounding background (Appendix A.6).

Mean

Mean signifies the mean value of all the tumor pixels (Appendix A.7).

Variance

Variance signifies the variance value of all the tumor pixels (Appendix A.8).

5.6.3 Histogram Features

A histogram in image processing tells about the frequency of each pixel's appearance in the image. A histogram is shown in figure 5.10.

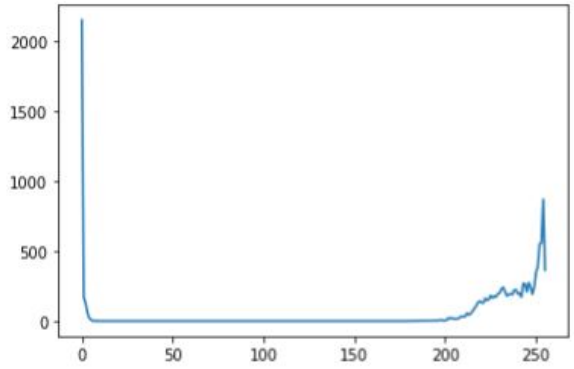


Figure 5.10: Histogram

The histogram features used along with their significance are listed below -

Histogram Mean

Histogram mean signifies the most commonly appearing pixel value in the image.

Histogram Variance

Histogram variance signifies the variance of all the frequencies in the histogram.

Histogram Peak

Histogram peak signifies the intensity value that occurs the most in the tumor region.

Histogram Skew

Histogram skew signifies how much the histogram varies from the normal bell-shaped curve.

5.7 Recommendation using TOPSIS

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) developed by Hwang and Yoon [9] is closely related to the idea that the recommendation should be such that the chosen alternative should have the least geographic distance from the positive-ideal solution and should be the farthest from negative-ideal solution. The positive-ideal solution maximizes the benefit criteria and minimizes the cost criteria. The negative-ideal solution maximizes the cost criteria and minimizes the benefit criteria.

The feature matrix obtained from the previous module is first normalised using min-max scaling (appendix A.9).

Standardization of matrix is required to ensure that we obtain dimensionless units as there may be different units of data measurements in the observations, which can be aggregated for rating and ranking the decision alternatives.

		CRITERIA										
		Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
ALTERNATIVES	Reimaging or Other	6	5	7	7	4	6	6	6	5	6	5
	Immediate Treatment	0	7	8	8	10	9	8	7	6	10	0
	No Treatment	10	1	2	2	1	2	1	2	1	2	10

Figure 5.11: Weight Matrix for TOPSIS

The next step involved in TOPSIS is to create a weight matrix. The weight matrix consists of the criterion and its weight on a predefined scale. The scale used in this project is 0-10. Higher the score, higher the dependency of an alternative on the criterion. The weight matrix used was obtained using a survey which was filled by a radiologist and reflects her observations.

The weight matrix is normalised using appendix A.10.

After obtaining the normalised weight matrix Euclidean distance to each alternative is calculated. Euclidean distance is the distance between two points in any plane or space as shown in appendix A.11.

After calculating the distance of the alternatives which in our case are no treatment, immediate treatment and re-imaging or other treatment methods, the

alternatives are ranked in accordance with the euclidean distance with each case or mammogram taken as the ideal solution. The alternatives with highest ranks gives the most probable treatment recommendations.

Algorithm 2: Recommendation using TOPSIS

Input: Feature matrix, weight matrix, number of features (**M**)
Result: Ranked recommendations
Procedure;
Normalise the feature matrix using min-max scalar;
Normalise the weight matrix by taking the ratio of each entry to the columnar root mean squared value;
for *each tumor feature* ***i*** *in normalised feature matrix* ***F*** **do**
  for *each alternative* ***j*** *in normalised wieght matrix* ***N*** **do**
    Find the Euclidean distance **d(j)** of the feature **F(i)** to alternative **N(j)**;
  end
  Sort **d**;
  for *each* ***x*** *in* ***M*** **do**
    Set **rank(i, x) = d(i)**
  end
end

5.7.1 TOPSIS: A Case

Step 1. Weight matrix shown in figure 5.11 is normalised using appendix A.10 (figure 5.12).

	CRITERIA										
ALTERNATIVES	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
Reimaging or Other	0.6	0.7	0.7	0.7	0.4	0.6	0.7	0.7	0.7	0.6	0.5
Immediate Treatment	0	0.86	0.72	0.72	0.92	0.79	0.77	0.71	0.73	0.82	0.89
No Treatment	0.84	0.11	0.18	0.18	0.09	0.17	0.09	0.2	0.12	0.16	0

Figure 5.12: Normalised Weight Matrix

A tuple from feature matrix from the normalised feature matrix is shown in figure 5.13.

Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
0.599662	0.544823	0.185483	0.884496	1	0.187019	0.448059	1	1	1	0.944063

Figure 5.13: A Tuple from Normalised Feature Matrix

Step 2. The tuple values are subtracted from the weight matrix values as shown in figure 5.14. The obtained values are shown in figure 5.15.

	CRITERIA										
ALTERNATIVES	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
Reimaging or Other	0.6 - 0.599662	0.7 - 0.54482	0.7 - 0.18548	0.7 - 0.88449	0.4 - 1	0.6 - 0.18701	0.7 - 0.44805	0.7 - 1	0.7 - 1	0.6 - 1	0.5 - 0.944063
Immediate Treatment	0 - 0.599662	0.86 - 0.54482	0.72 - 0.18548	0.72 - 0.88449	0.92 - 1	0.79 - 0.18701	0.77 - 0.44805	0.71 - 1	0.73 - 1	0.82 - 1	0.89 - 0.944063
No Treatment	0.84 - 0.599662	0.11 - 0.54482	0.18 - 0.18548	0.18 - 0.88449	0.09 - 1	0.17 - 0.18701	0.09 - 0.44805	0.2 - 1	0.12 - 1	0.16 - 1	0 - 0.944063

Figure 5.14: Subtraction of Values

	CRITERIA										
ALTERNATIVES	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
Reimaging or Other	0.000338	0.15518	0.51452	-0.1845	-0.6	0.41298	0.25194	-0.3	-0.3	-0.4	-0.44406
Immediate Treatment	-0.599662	0.31518	0.53452	-0.1645	-0.08	0.60298	0.32194	-0.29	-0.27	-0.18	-0.05406
No Treatment	0.240338	-0.4348	-0.0055	-0.7045	-0.91	-0.017	-0.3581	-0.8	-0.88	-0.84	-0.94406

Figure 5.15: Subtraction of Values

Step 3. The values obtained in step 2 are squared as shown in figure 5.16.

Step 4. The sum of values (obtained in step 2) is taken for each alternative as shown in figure 5.17.

Step 5. To obtain the Euclidean distance as per appendix A.11, square root is taken for the sum and ranks are assigned in accordance with ascending values, shown in figure 5.18.

ALTERNATIVES	CRITERIA											
	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew
Reimaging or Other	1.14E-07	0.02408	0.26473	0.03404	0.36	0.17055	0.06347	0.09	0.09	0.16	0.197192
Immediate Treatment	0.359595	0.09934	0.28571	0.02706	0.0064	0.36359	0.10365	0.0841	0.0729	0.0324	0.002923
No Treatment	0.057762	0.18907	3E-05	0.49631	0.8281	0.00029	0.12821	0.64	0.7744	0.7056	0.891255

Figure 5.16: Squared Values

ALTERNATIVES	CRITERIA											
	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew	SUM
Reimaging or Other	0.0000001	0.02408	0.26473	0.03404	0.36	0.17055	0.06347	0.09	0.09	0.16	0.197192	1.45407
Immediate Treatment	0.3595945	0.09934	0.28571	0.02706	0.0064	0.36359	0.10365	0.0841	0.0729	0.0324	0.002923	1.43765
No Treatment	0.0577624	0.18907	0.00003	0.49631	0.8281	0.00029	0.12821	0.64	0.7744	0.7056	0.891255	4.71103

Figure 5.17: Sum of Values for each Alternative

ALTERNATIVES	CRITERIA												
	Roundness	Acreage Ratio	Energy	Entropy	Contrast Coefficient	Texture Mean	Texture Variance	Histogram Mean	Histogram Variance	Histogram Peak	Histogram Skew	SQRT SUM	RANK
Reimaging or Other	0.0000001	0.02408	0.26473	0.03404	0.36	0.17055	0.06347	0.09	0.09	0.16	0.197192	1.20585	2
Immediate Treatment	0.3595945	0.09934	0.28571	0.02706	0.0064	0.36359	0.10365	0.0841	0.0729	0.0324	0.002923	1.19902	1
No Treatment	0.0577624	0.18907	0.00003	0.49631	0.8281	0.00029	0.12821	0.64	0.7744	0.7056	0.891255	2.17049	3

Figure 5.18: Distance with Corresponding Ranks

5.8 Project Roadmap

5.8.1 Stage 1: Literature Survey

Various papers with respect to all the processes and modules in this research were referred and a base model to approach the project was developed.

5.8.2 Stage 2: Planning

Then the initial idea was discussed with the guides namely, Dr Sanjay Pawar and Dr Debabrata Datta. The modules to be implemented were studied and algorithms concerning the same were chosen.

5.8.3 Stage 3: Requirement Gathering

The mammogram data was obtained, platform and environment for the project was chosen. All the required packages were downloaded the in the environment.

5.8.4 Stage 4: Implementation

The data was studied and then preprocessed according to the system requirements. Then the image processing of pectoral muscle, CNN, ROI extraction and feature extraction modules were implemented.

5.8.5 Stage 5: A Visit to Tata Memorial Centre, Mumbai

A survey of weights for the recommendation and ROI detection modules was conducted. She also provided her insights on the features used and suggested some others which could also be used. After securing weights the recommendation module was implemented.

5.8.6 Stage 6: Validation and Fine Tuning

For validating the results suspicious mammograms were independently examined by Dr Asawari and hence the recommendations were verified. For fine tuning of the CNN module, hyper-parameters like learning rate, decay, optimiser functions were tuned and the results were noted.

Chapter 6

Results and Discussion

6.1 Image Processing of Pectoral Muscle

The accuracy obtained using algorithm 1 (as opposed to the linear cut-off method [6]) in classification of mammogram as normal and abnormal in case of dense tissues using random forest classifier in percentage is shown below.

Linear Cut-off	Proposed System
79.16	83.30

6.2 Detection

6.2.1 Summary of CNN

The layers of the convolutional neural network designed is summarised in figure 6.1. Since the problem in discussion is regression, exact pin-pointing of the point is tedious, any value of accuracy close to 70% is a good result as is shown in figures 6.4, 6.5 and 6.6. The metric and loss function for the same are accuracy and mean squared error. Since the dataset used is relatively small, k-fold cross validation is used for best results.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
batch_normalization (BatchNo (None, 512, 512, 1)       4
_________________________________________________________________
conv2d (Conv2D)              (None, 508, 508, 24)      624
_________________________________________________________________
activation (Activation)      (None, 508, 508, 24)      0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 254, 254, 24)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 250, 250, 36)      21636
_________________________________________________________________
activation_1 (Activation)    (None, 250, 250, 36)      0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 125, 125, 36)      0
_________________________________________________________________
flatten (Flatten)            (None, 562500)            0
_________________________________________________________________
dense (Dense)                (None, 3)                 1687503
=================================================================
Total params: 1,709,767
Trainable params: 1,709,765
Non-trainable params: 2
_________________________________________________________________
```

Figure 6.1: CNN Summary

6.2.2 Optimisers

A total of five optimisers were tried out, the result of each one is given below -

ADADELTA

ADADELTA is a per-dimension learning rate method for gradient descent. There is no need of manually setting a learning rate and is invariant to noisy gradient, model architecture, data modalities and hyper-parameters.

Adam

Adam has relatively low memory requirements and works with little tuning of hyper-parameters.

Adamax

Adamax is a variant of Adam optimiser, it works on infinity norm.

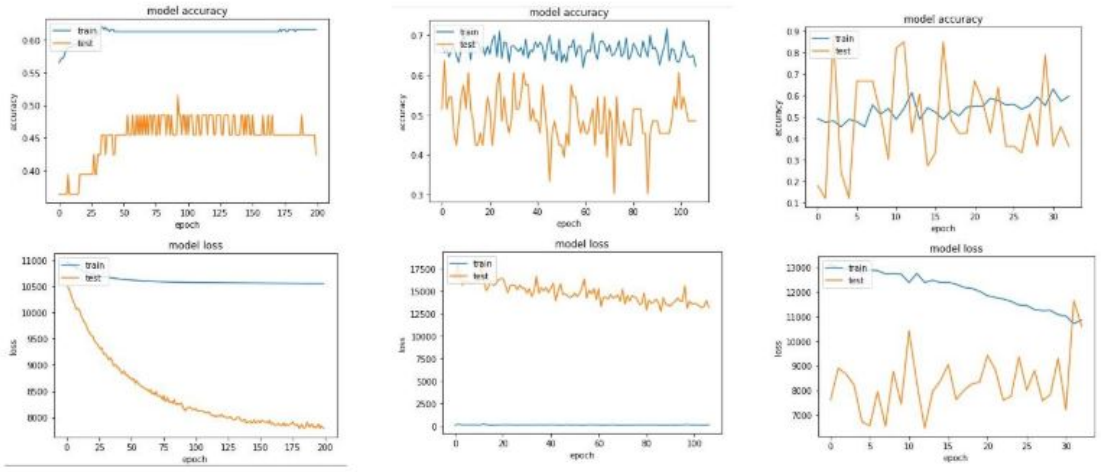


Figure 6.2: Adadelta (L), Adam (C), Adamax (R)

RMSProp

RMSprop can be viewed as a way to deal with its radically diminishing learning rates.

Nadam

Nadam is Adam with Nesterov momentum.

Adagrad

Adagrad provides with parameter-specific learning rates, which adapt relative to how often a parameter gets updated during training. The more the updates, the smaller the learning rate.

Since RMSProp had gradual decrease in loss and steady accuracy both in validation as well as training, it was chosen as the ideal model.

The result obtained after thorough training of the model is given below -

- Training accuracy - **85.43%**

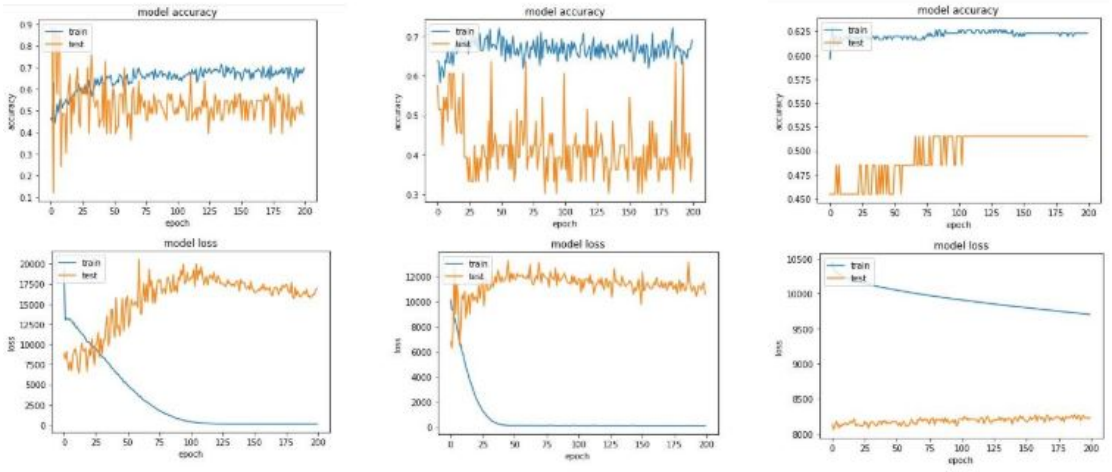


Figure 6.3: RMSProp (L), Nadam (C), Adagrad (R)

- Testing accuracy - **73.49%**

6.3 Samples From the Testing Set

A few of the samples are shown in the following figures to emphasise on the result.

Cases with No Abnormality

The model was trained in such a way that whenever it came across any sample where there is no abnormality, it returned centre coordinates as (1024, 1024) and radius as 0. The region marked by the algorithm is shown in blue color. Figure 6.4 shows a few samples from the testing dataset.

Cases with Tumor Present

The actual region is marked by red color whereas the region predicted is marked by blue color. Figures 6.5 and 6.6 show a few samples from the testing dataset.

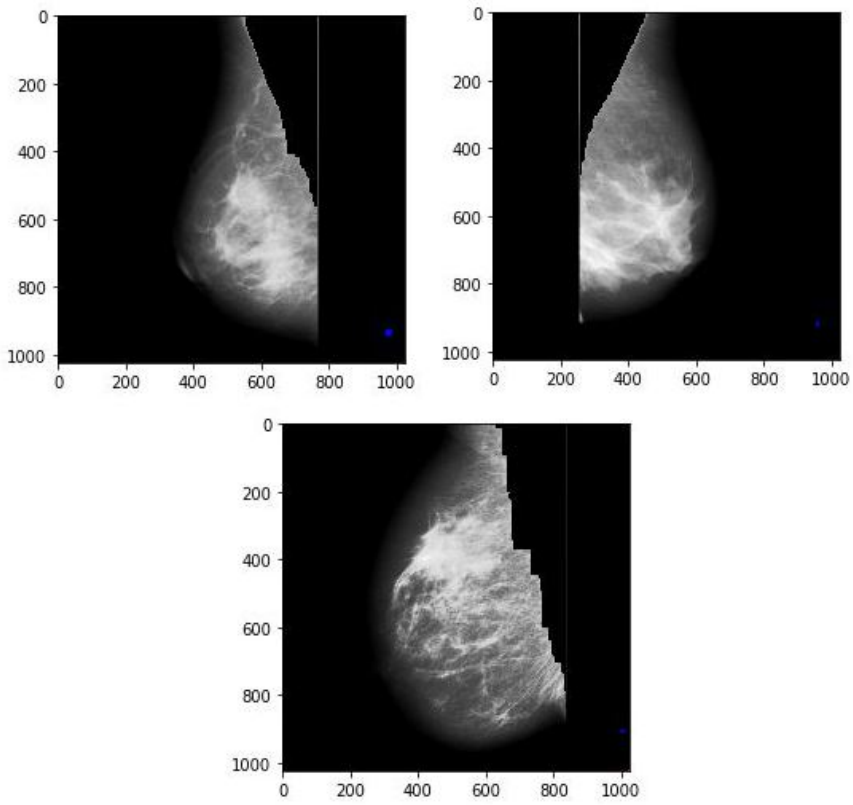


Figure 6.4: No Abnormality Cases

6.4 Recommendation

A total of 119 suspicious tumors were detected using the detection module. Features as mentioned in the earlier chapters were extracted from each of the tumors. These features were closely evaluated by a radiologist and scores were assigned to each feature with respect to the type of treatment. Since, the original dataset did not already contain the recommendation labels, the 119 cases were independently evaluated by a radiologist.

Since TOPSIS is a ranking algorithm, the accuracy is measured by checking if the correct recommendation is made in the first two ranks. With a sensitive application like that of cancer treatment recommendation, merely relying only on the first rank may seem incomplete as the patient may also be in need of a second choice.

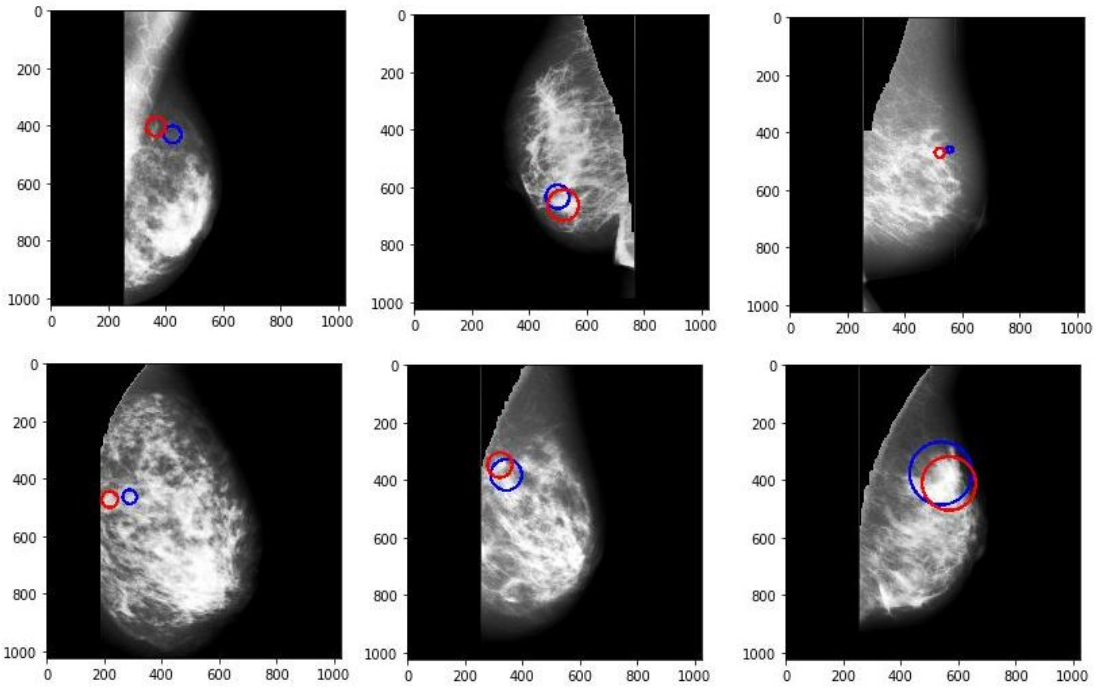


Figure 6.5: Abnormal Cases 1

The algorithm designed recommends the correct treatment within the first two ranks **97 out of 119** times leading to an accuracy of 81.5%.

6.4.1 Significance of Alternatives

Re-imaging or Other Methods

Recommendation of this alternative may be attributed to the following -

- A deeper understanding of the case may be required by the means of Magnetic Resonance Imaging (MRI) or Ultra Sonography (USG)
- The breast captured may have a high tissue density
- The image may not be clear

Immediate Treatment

Recommendation of this alternative signifies -

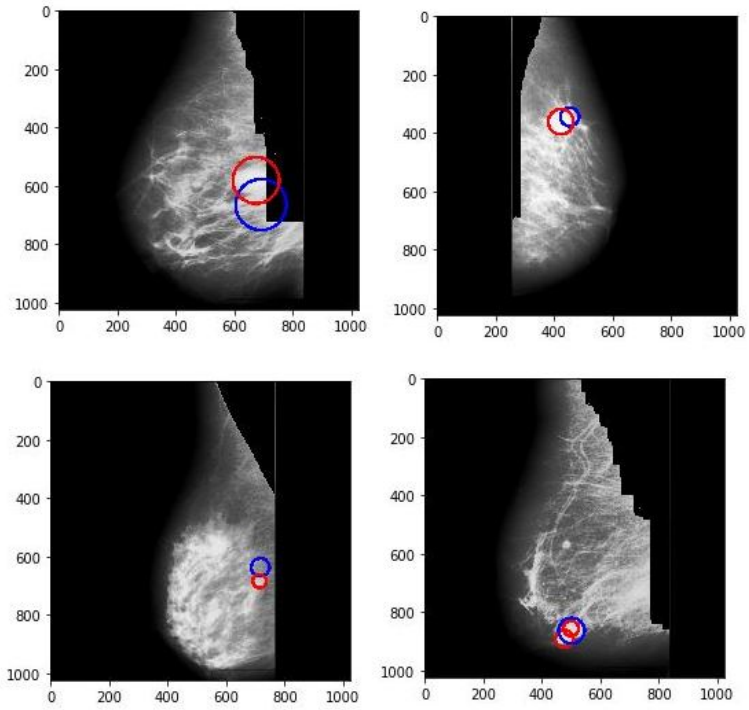


Figure 6.6: Abnormal Cases 2

- The case may be of utmost seriousness (malign)
- The case may be benign but needs to be addressed

No Treatment

Recommendation of this alternative may be because the identified tumor region can go without being addressed.

However, it is worth noting that the results obtained is solely based on the dataset processed and there may be a need of adding more alternatives for addressing more serious cases.

Chapter 7

Conclusion

With times and situations where visiting a doctor may not always be possible, it is desirable to have an application that can make diagnosis and treatment recommendation easier for patients. Our efforts can be extended to a product which will make it more accessible to the women of all generations. This standalone product will not only give an idea about the tumor but also will guide the patient towards the most suited treatment. The following sections discuss the advantages, disadvantages, applications and future scope.

7.1 Advantages

Following are some of the features which sets the system proposed apart from the other similar CAD models.

- The system is highly modular and each module is cohesive
- Along with recommendations, the system also returns eleven associated features about the abnormality
- Features used for recommendation are such which doctors actually consider while evaluating a case
- The system leaves a wide window for adding new criteria (features) and alternatives (treatments) thus making the system highly flexible

- Weights for TOPSIS can be modified with respect to the case in consideration - age for instance, younger female breasts have lower tissue density as compared to an adult
- Mammography is the most common method of imaging, hence the system will have a broad reach
- With appropriate weights, the system can be easily extended for other types of cancer

7.2 Disadvantages

While the system built attains a good accuracy in terms of detection and recommendation, there are a few limitations that could be identified in the system. The outcome can be further refined by using a bigger dataset.

- The system built solely processes mammogram images and does not factor other imaging methods in
- The mammogram to be used must be the best version of itself
- The system may tend to be sensitive towards high tissue density
- A good precision is required while processing pectoral muscle so as to not accidentally remove any of the breast mass
- Weights for the recommendation module are fixed and not set as learning quantities

7.3 Applications and Future Scope

Such a system can be easily integrated into a web application that will enable early detection of abnormality. In absence of a specialist, this system can give an insight about the grave seriousness of the issue in hand. It may also be deployed in pathological laboratories where mammography imaging is done, so that right

after getting a procedure done, the patient be informed about the abnormality, understand how severe it is and what should be done next.

When radiologists actually examine a case, various other factors along with the ones considered here are analysed - like age, blood group, height, weight, medical records, marginal adhesion, enlarged epithelial cells, bare nuclei, etc. or even other imaging techniques like MRI, USG, biopsy, histopathology, etc. The system can be extended to include more features or imaging techniques. The system built as of now only deals with breast abnormalities and hence could easily be extended to other types of cancer as well.

Appendices

Appendix A

Mathematical Equations

A.1 Roundness

$$r = \frac{p^2}{4\pi A}$$

A.2 Acreage Ratio

$$a = \frac{Area(Contour)}{Area(MEC)}$$

A.3 Inverse Moment

$$I = \sum \frac{P(i,j)}{1 + (i-j)^2}$$

A.4 Energy

$$E = \sum P(i,j)^2$$

A.5 Entropy

$$S = \sum P(i,j) \cdot (-ln(P(i,j)))$$

A.6 Contrast Coefficient

$$C = \sum (i-j)^2 \cdot P(i,j)$$

A.7 Mean

$$\mu = \frac{1}{N} \sum P(i,j)$$

A.8 Variance

$$\sigma^2 = \sum \frac{(P(i,j) - \mu)^2}{N}$$

A.9 Min-max Scaling

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

A.10 Normalising Weight Matrix

$$X_{inorm} = \frac{X_i}{\sqrt{\sum x_i^2}}$$

A.11 Euclidean Distance

$$d = \sqrt{\sum (x_i - y_i)^2}$$

Appendix B

Questionnaire

Dr Asawari Lautre, radiologist at Tata Medical Centre, Mumbai, was asked to rate the obtained features on the basis of its importance in practice. She was also asked to mark the tumor region out of the extracted abnormal region. Her valuable suggestions were noted and appropriate changes were made accordingly.

The filled survey form is shown below -

Breast Tumor Evaluation Questionnaire

Please indicate the importance of the following features extracted from the mammogram images available on a scale of 0 to 10 for recommendation of respective probable treatments –

1. **Roundness**
   Significance – how much similar the shape of the tumor is to an ideal circle. The metric typically lies between **0 and 1**.
   0 - not a circle 1 – perfect circle

	Reimaging required	Immediate treatment required	No treatment required
Score	6-7.	0	10.

2. **Acreage ratio (land to building ratio)**
   Significance – how much area of an ideal circle does the tumor occupy

	Reimaging required	Immediate treatment required	No treatment required
Score	5-6.	7	1.

3. **Energy**
   Significance – how much the neighboring pixels vary from each other

	Reimaging required	Immediate treatment required	No treatment required
Score	7-8.	8	2

4. **Entropy**
   Significance – how irregular the texture of the tumor is

	Reimaging required	Immediate treatment required	No treatment required
Score	7-8.	8	2

5. **Contrast coefficient**
   Significance – how bright is the tumor from it surrounding background

	Reimaging required	Immediate treatment required	No treatment required
Score	4.	10	1.

6. **Inverse moment**
   Significance –

	Reimaging required	Immediate treatment required	No treatment required
Score			

7. **Mean**
   Significance – the mean value of all the tumor pixels

	Reimaging required	Immediate treatment required	No treatment required
Score	6-7.	9	2

8. **Variance**
   Significance – the variance value of all the tumor pixels

	Reimaging required	Immediate treatment required	No treatment required
Score	6-7.	8	1

A histogram is typically a value vs count chart i.e. 255 intensity values vs the number of pixels of each intensity.

9. **Histogram mean**
   Significance – the most common occurring intensity value

	Reimaging required	Immediate treatment required	No treatment required
Score	6-7.	7	2.

10. Histogram variance

Significance – the variance of all the intensity values

	Reimaging required	Immediate treatment required	No treatment required
Score	5-6.	6.	1.

11. Histogram peak

Significance – the intensity value that occurs the most in the tumor region

	Reimaging required	Immediate treatment required	No treatment required
Score	6-7.	10.	2.

12. Histogram skew

Significance – how much the histogram varies from the normal distribution

	Reimaging required	Immediate treatment required	No treatment required
Score	—	Right side skewed.	Left to centre.

Are there any other significant features we are missing out on?

Yes, margins of the tumor, breast density in the background and also age.

What are the other existing methods other than mammography is used for detection?

USG, MRI, Biopsy.

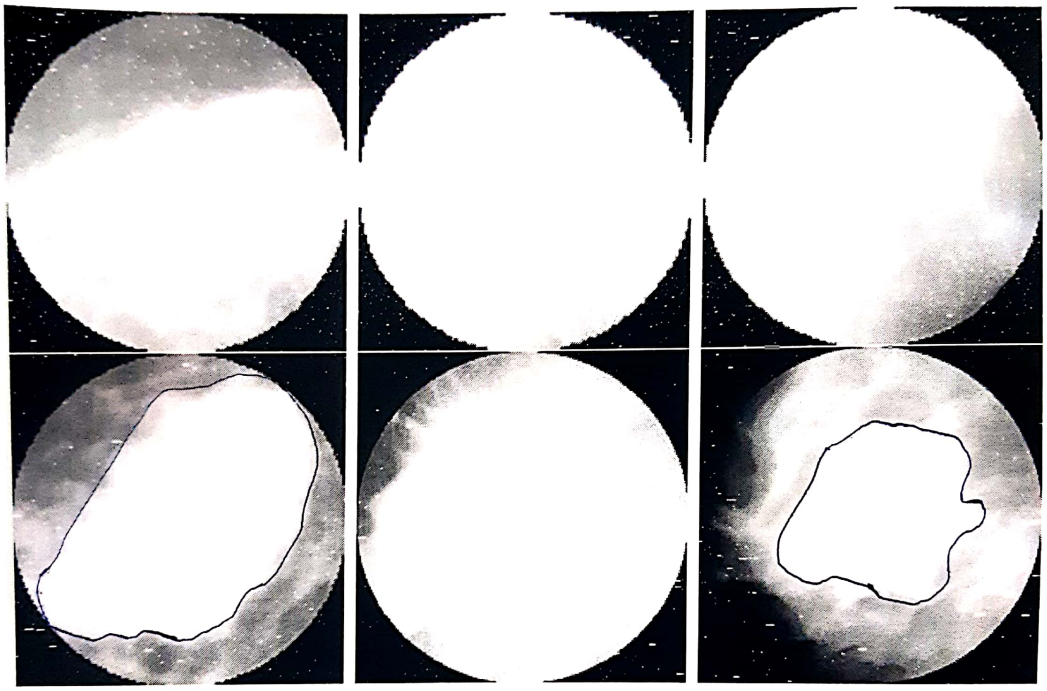
Will a system like the one which we intend to build be helpful for early detection?

Yes.

Any other suggestions

Better take plain MG than contrast enhanced MG.

Please indicate the tumor in the regions shown below. This will help us to set a threshold that automatically identifies the tumor.



Name – Dr. Asawari S. Lautre

Date – 13 / 3 / 2020

asawarilautre24@gmail.com

Signature

Appendix C

Important Concepts

C.1 Recommendation Systems

Various types of recommendation techniques exists other than TOPSIS. Recommender systems help suggesting relevant items to users and are algorithm based. There are two approaches.

C.1.1 Collaborative Filtering Approach

Collaborative methods for recommender systems are based on the past interactions recorded between users and items to produce new recommendations to users. These interactions are stored in the user-item interactions matrix. These past user-item interactions are used to detect similar users and/or similar items and make predictions based on these estimated proximities.

The collaborative filtering could be model based or memory based.

Model Based

A model for user-items interaction is defined where users and items representations have to be learned from interactions matrix.

Memory Based

Define no model for user-item interaction and rely on similarities between users

or items based on their interactions.

The main advantage of collaborative is that they do not acquire information about users or items which makes them useful in many situations. Also, the more users interact with items the more new recommendations become accurate.

However, as it only considers past interactions to make recommendations, it is impossible to recommend anything to new users and many users or items have too few interactions to be efficiently handled which is known as a cold start problem. This drawback can be overcome by recommending random items to new users or new items to random users (random strategy), recommending popular items to new users or new items to most active users (maximum expectation strategy), recommending a set of various items to new users or a new item to a set of various users (exploratory strategy) or using a non-collaborative method altogether.

C.1.2 Content Based Filtering Approach

In content based methods, the recommendation problem is casted into either a classification problem or into a regression problem. In both cases, the model will be based on the user and/or item features at our disposal.

If the classification or regression is based on users features then it is an item-centered approach. In this case, model is build and learned by item based on users features trying to answer the question "what is the probability for each user to like this item?". The model associated to each item is naturally trained on data related to this item and it leads, in general, to pretty robust models as a lot of users have interacted with the item. However, the interactions considered to learn the model come from every users and even if these users have similar characteristic features then their preferences can be different.

If we are working with items features, the method is then user-centered. Then the model is trained based on items features that tries to answer the question "what is the probability for this user to like each item?". A model to each user can be attached that is trained on its data. The model obtained is therefore more personalised than its item-centered counterpart because it only takes into account interactions from the user into consideration. However, most of the time a user has interacted with relatively few items and therefore the model obtained is a far

less robust than an item-centred one.

C.1.3 Hybrid Approach

This approach combines collaborative filtering and content based approaches, achieves state-of-the-art results in many cases and is therefore used in many large scale recommendation systems nowadays. There are two types for combination made in hybrid approaches .

The two models can be trained independently (one collaborative filtering model and one content based model) and combine their suggestions or directly build a single model (often a neural network) that unify both approaches by using as inputs prior information as well as "collaborative" interactions information.

C.2 Neutrosophic Set Theory

Neutrosophy is a field of philosophy which studies the origin, nature, and scope of neutralities, as well as their interactions with different ideational spectra.

Etymologically, neutro-sophy in French neutre or in Latin neuter means neutral and in Greek Sophia means skill or wisdom therefore neutron-sophy means knowledge of neutral thought and was started in 1995.

The Fundamental theory states that Every idea <A> tends to be neutralized, diminished, balanced by ¡nonA¿ ideas as a state of equilibrium.

<nonA> = what is not <A>,

<antiA> = the opposite of <A>, and

<neutA> = what is neither <A> nor <antiA>.

In a classical way <A>, <neutA>, <antiA> are disjoint two by two but, since in many cases the borders between notions are vague and imprecise it is possible that <A>, <neutA>, <antiA> have common parts two by two as well.

This could be also explained In the following manner

Consider the nonstandard unit interval [-0, 1+] with imprecise left and right borders. Let T, I, F be standard or nonstandard subsets of the interval [-0, 1+]. Neutrosophic Logic is a logic in which each proposition is T% true, I% indeterminate, and F% false.

$-0 \leq \inf T + \inf I + \inf F \leq \sup T + \sup I + \sup F \; leq \; 3+$

T, I, F are not necessary intervals, but any sets which could be discrete or

continuous, open or closed or half- open or half-closed interval, intersections or unions of the previous sets.

For example Proposition P is between 30-40% can be also 45-50% true, 20% indeterminate, and 60% or between 66-70% false according to various parameters. Neutrosophic Logic is a generalization of Zadeh's fuzzy logic (FL), and especially of Atanassov's intuitionistic fuzzy logic (IFL), and of other logics.

References

[1] Mohammed A Al-masni, Mugahed A Al-antari, JM Park, Geon Gi, Tae-Yeon Kim, Patricio Rivera, Edwin Valarezo, S-M Han, and T-S Kim. Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1230–1233. IEEE, 2017.

[2] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensarİ. Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4. IEEE, 2018.

[3] Anu Appukuttan and L Sindhu. Curvelet and pnn classifier based approach for early detection and classification of breast cancer in digital mammograms. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 1, pages 1–5. IEEE, 2016.

[4] Ebru Aydındag Bayrak, Pınar Kırcı, and Tolga Ensari. Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–3. IEEE, 2019.

[5] Yu-Tso Chen, Wen-Chun Peng, and Hsin-Yu Yu. Identify key factors for career choice by using topsis and fuzzy cognitive map. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 104–109. IEEE, 2018.

[6] M. Hanmandlu, A. A. Khan, and A. Saha. A novel algorithm for pectoral muscle removal and auto-cropping of neoplasmic area from mammograms. In *2012 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5, 2012.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Kai Hu, Xieping Gao, and Fei Li. Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms. *IEEE Transactions on Instrumentation and Measurement*, 60(2):462–472, 2010.

[9] Ching-Lai Hwang and Kwangsun Yoon. Methods for multiple attribute decision making. In *Multiple attribute decision making*, pages 58–191. Springer, 1981.

[10] Nur Syahmi Ismail and Cheab Sovuthy. Breast cancer detection based on deep learning.

[11] Varsha T Lokare and Prakash M Jadhav. Using the ahp and topsis methods for decision making in best course selection after hsc. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE, 2016.

[12] Hao-Chun Lu, El-Wui Loh, and Shih-Chen Huang. The classification of mammogram using convolutional neural network with specific image preprocessing for breast cancer detection. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 9–12. IEEE, 2019.

[13] Nada A Nabeeh, Florentin Smarandache, Mohamed Abdel-Basset, Haitham A El-Ghareeb, and Ahmed Aboelfetouh. An integrated neutrosophic-topsis approach and its application to personnel selection: A new trend in brain processing and analysis. *IEEE Access*, 7:29734–29744, 2019.

[14] Robert M Nishikawa. Computer-aided detection and diagnosis. In *Digital Mammography*, pages 85–106. Springer, 2010.

[15] University of South Florida. University of south florida, digital mammography, home page.

[16] World Health Organisation. International agency for research in cancer.

[17] Samuel Rahimeto, Taye Girma Debelee, Dereje Yohannes, and Friedhelm Schwenker. Automatic pectoral muscle removal in mammograms. *Evolving Systems*, pages 1–8, 2019.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] Mohammad Taheri, George Hamer, Seong Ho Son, and Sung Y Shin. Enhanced breast cancer classification with automatic thresholding using svm and harris corner detection. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pages 56–60, 2016.

[20] Zhiqiong Wang, Mo Li, Huaxia Wang, Hanyu Jiang, Yudong Yao, Hao Zhang, and Junchang Xin. Breast cancer detection using extreme learning machine based on feature fusion with cnn deep features. *IEEE Access*, 7:105146–105158, 2019.

[21] Wikipedia. Mammography.

[22] Kewen Yan, Shaohui Huang, Yaoxian Song, Wei Liu, and Neng Fan. Face recognition based on convolution neural network. In *2017 36th Chinese Control Conference (CCC)*, pages 4077–4081. IEEE, 2017.