Chapter 1

INTRODUCTION

1.1 Introduction

Breast cancer is the most common cancer of American women and is the leading cause of cancer-related death among women aged 15-54. In 1996, the American Cancer Society estimated that 184,300 women will be diagnosed with breast cancer and that 44,300 women will die from it. Another study showed approximately 720,000 new cases will be diagnosed world-wide per year, this accounts for about 20% of all malignant tumour cases. Early detection and diagnosis of breast cancer is one of the most important factors affecting the possibility of recovery from the disease.

The World Health Organization's International Agency for Research on Cancer estimates that more than 150,000 women worldwide die of breast cancer each year. Since breast cancer is a progressive disease, evolving through stages of cellular dedifferentiation and growth, the time at which breast cancer is detected is crucial. The earlier breast cancer is detected, the higher is the chance of survival.

Screening Mammography is the only method currently available for the reliable detection of early and potentially curable breast cancer. In mammography, the low energy X-rays are used for creating and examining human breast and can help in cancer detection at the earliest stage, by detecting small calcium deposits. Mammography is high-resolution x-ray imaging of the compressed breast. This involves radiation transmission through the tissue and the projection of anatomical structures on a film screen or image sensor. Associated with the x-ray imaging projection is a reduction in anatomical information from a 3D organ to a 2D film/image.

Two imaging projections of each breast, Craniocaudal (CC) and mediolateral oblique (MLO) views are routinely obtained. This permits some indication of three dimensions and an understanding of overlapping structures. High quality mammogram with high spatial resolution and adequate contrast separation allows

radiologists to observe fine structures. Studies have shown that the mortality 2 Mediolateral Oblique (MLO) view Craniocaudal (CC) view.

Mammography is used to detect the Malignant or Benign tumors based on the differential absorption of X-rays between the various tissue components of the breast such as fat, fibroglandular tissue, extremely dense tissue (can be a tumor) and calcifications.

1.2 Motivation

With the wide spread development of screening programs in the USA, radiologists have had to read a large number of mammograms. Reading mammograms is difficult and requires a great deal of experience. Several studies have shown retrospectively that 20% to 40% of breast cancer fails to be detected at screening due to radiologist fatigue, the complex image structure of the breast tissue, and the subtlety of the cancer. Even the most experienced mammographic readers only have a correct detection rate of 85-91%.

Moreover, a study found that there is about 2.6% to 15.9% false positive reading of negative or benign mammograms by radiologists. Several studies showed that double reading by two radiologists can improve detection sensitivity up to 15%.

However, implementing double reading can be very costly, time consuming and logistically difficult. It has been proposed that a computer aided diagnostic (CAD) system be used as a second reader to assist the radiologist, leaving the final decision to the human.

CAD can increase the diagnostic accuracy and efficiency with high reproducibility. It has shown that the performance of a radiologist can be increased 5-15% by providing the radiologist with results from a CAD system as a “second opinion”. It has also been shown that a CAD system can detect approximately 50% of the lesions which are missed at screening.

1.3 Image Classification

Image classification refers to the task of extracting information classes from a multiband raster image. The resulting raster from image classification can be used to create thematic maps. Depending on the interaction between the analyst and the computer during classification, there are two types of classification: supervised and unsupervised.

1.3.1 Supervised classification

Supervised classification uses the spectral signatures obtained from training samples to classify an image. With the assistance of the Image Classification toolbar, you can easily create training samples to represent the classes you want to extract. You can also easily create a signature file from the training samples, which is then used by the multivariate classification tools to classify the image.

1.3.2 Unsupervised classification

Unsupervised classification finds spectral classes (or clusters) in a multiband image without the analyst's intervention. The Image Classification toolbar aids in unsupervised classification by providing access to the tools to create the clusters, capability to analyse the quality of the clusters, and access to classification tools.

1.4 Problem Statement

To classify the given Mammogram into benign or malignant based on the deposition of microcalification in the Mammogram and to detect the condition of background tissue and the state of abnormality. After the classification of Mammogram Images into Benign or Malignant various parameters are verified and checked which will predict the stage of Cancer.

1.4.1 Stages of breast cancer:

Stage 0: Signs of cancer cells. (Earlier Stage)
Stage 1: The cancer cells are contained to very limited area. (Earlier Stage)
Stage 2: Cancer has begun to grow and spread. (Advance Stage)
Stage 3: Cancer is invading tissue around the breast. (Advance Stage)
Stage 4: Cancer has spread beyond the breast to other body.

1.5 Description of the Dataset

1.5.1 Dataset: Mini-MIAS

The Mammographic Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. Films taken from the UK National Breast Screening Programme have been digitised to 50-micron pixel edge with a Joyce-Loebl scanning microdensitometer, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains 322 digitised films and is available on 2.3GB 8mm (ExaByte) tape. It also includes radiologist's "truth"-markings on the locations of any abnormalities that may be present. The database has been reduced to a 200-micron pixel edge and padded/clipped so that all the images are 1024x1024. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex.

Mini-MIAS database which contains 322 digitized mammogram images consisted of left and right breast images. The acquired mammogram images are classified into three distinct classes: Fatty (F) (106 images), Fatty-Glandular (G) (104 images) and Dense-Glandular (D) (112 images). Size of these images is 1024 × 1024 pixels in Portable Grey map (PGM) format. Each pixel in the images is represented as an 8-bit word, where the images are in grayscale format with a pixel intensity of range from 0 to 255.

1.6 Summary

Breast cancer is one of the commonest types of cancer contributing to the increase in mortality among women worldwide. Recent statistics show that breast cancer affects one of every ten women in Europe and one of every eight in the United States.

It is important to accurately diagnose Mammograms to detect the Cancer at the early stages. Early detection and diagnosis of breast cancer is one of the most important factors affecting the possibility of recovery from the disease. CAD (Computer Aided Diagnosis) can be employed as supplement to the Radiologist’s assessment.Therefore, reducing the False Negative and False Positive cases.

Chapter 2

LITERATURE SURVEY

2.1 Overview

The main hindrance in processing medical images such as X-rays, MRIs, Mammograms and other angiographic images is that images get corrupted by noise during their acquisition and transmission. For additive noise removal, Image Denoising techniques are applied. Image Denoising is process of removal of noise in the Images. Mostly, statistical filters like Median Filter, Wiener Filter and Gaussian Filters are applied to remove the noise from the Image.

But using those statistical filters resulting in blurred out edges and meaningful textures in Images. As Medical Images relies on the Edges and Textures for diagnosing, these statistical filters are not suitable for CAD (Computer Aided Diagnosis) systems. Once the image is denoised it can be segmented using algorithms like k means algorithm and after that feature is extracted on various factors and then the mammograms are classified.

2.2 Pre-Processing:

E. Malar, A. Kandaswamy, SS. Kirthana and D. Nivedhitha proposed several techniques for Image denoising in Medical Images in their "**A comparative study on mammographic image denoising technique using wavelet, curvelet and contourlet transforms**" [1]. They proposed that Transformation techniques such as Wavelet, Curvelet and Contourlet transforms for Image Denoising in Medical Images.

A good denoising technique should remove the overall noise in Image while maintaining the image boundaries, object edges and without spoiling the texture of the Image. Curvelet, Wavelet and Contourlet transforms are suitable for medical images as they tend to remove noise in the Image while maintaining the Image boundaries, object edges without spoiling the texture of the Image.

Various noises like Poisson, Gaussian, Speckle and Salt & pepper noises are considered to compare the Wavelet, Curvelet and Contourlet transforms. These

transformation performances are measured in terms of Peak Signal to Noise ratio (PSNR).

Wavelet transform can achieve good sparsity localized details, such as edges and singularities. For typical natural images, most of the wavelet coefficients have small magnitudes, except for a large one that represent that represent important high-frequency features of the image such as edges. The DWT (Discrete Wavelet Transform) is used for denoising the Image.

Curvelet transform is a multiscale geometric wavelet transforms, can represent edges and curve-singularities much more efficiently than traditional wavelet. Curvelet combines rate of converges by simple thresholding. Multiscale decomposition links point discontinuities into linear structures.

Where in, Contourlet transform is a discrete extension of Cuvelet transform that aims to capture the intrinsic geometrical structures of images and curves instead of points and provides directionality and anisotropy.

It can be observed from the table 1 that the curvelet method gives the best results for most of the noises comparatively to all other methods in terms of the SNR values. For Poisson noise contourlet and wavelet thresholding seem to produce better results than curvelet denoising. But, the slender substantial difference in the SNR results for those noises is comparatively inconsequential. Further, as the noise in the image is increased, curvelet based algorithms show significant improvement than other methods. The below table shows the result.

Table 2.1: Comparison of Wavelet, Contourlet and Curvelet based Denoising Methods

Noises	Wavelet SNR/dB	Contourlet SNR/dB	Curvelet SNR/dB
Salt & pepper	6.76	11.65	32.94
Poisson	25.71	25.55	23.1428
Gaussian	16.96	17.84	57.51
Speckle	16.69	19.13	22.1257

Contrast enhancement is the next step following denoising and is essential as mammograms have very low contrast and brightness preservation is also necessary. Enhancement methods can be classified broadly into direct and in-direct. Different kinds of indirect methods include, Histogram Equalization (HE), Recursive Mean

Separate Histogram Equalization (RMSHE), Contrast Limited Adaptive Histogram Equalization (CLAHE) etc.

Nabin Kharel, Abber Alsadoon, P.W.C Prasad, A. Elchouemi have compares several Contrast enhancement techniques along with proposing a new method in "**Early Diagnosis of Breast Cancer Using Contrast Limited Adaptive Histogram Equalization(CLAHE) and Morphology Methods**" [2].

In this work they compared Histogram Equalization (HE), Brightness Preserving Bi-histogram Equalization(BBHE), Dualistic Sub-Image Histogram Equalization (DSIHE), Minimum Mean Brightness Error Bi-Histogram Equalization (MMBEBHE) and their own method. The proposed method outperformed all other Histogram Equalization Techniques.

A good contrast enhancement technique should increase the contrast in the image without meddling with the Brightness of the image. These classic HE methods are too absorbed in preserving the input brightness. If the input image looks like dark, the enhancement effect of these methods may be not good because the output image also looks like dark.

Histogram equalization transforms the histogram of the original image into a flat uniform histogram with a mean value that is in the middle of gray level range. Accordingly, the mean brightness of the output image is always at the middle or close to it in the case of discrete implementation regardless of the mean of the input image. For images with high and low mean brightness values, this means a significant change in the image outlook for the price of enhancing the contrast. To overcome such drawbacks, variations of the classic HE technique such as the mean preserving bi-histogram equalization (BBHE) equal area dualistic sub-image histogram equalization (DSIHE) and minimum mean brightness error bi-histogram equalization (MMBEBHE) have been proposed.

Although the above methods have proved to be better but they introduce some unnatural noise in the image. To overcome this new technique called Contrast Limited Adaptive Histogram Equalization (CLAHE) has been proposed.

Adaptive histogram equalization divides the Image into small blocks called "tiles" (tileSize is 8x8 by default in OpenCV). Then each of these blocks are histogram equalized as usual. So, in a small area, histogram would confine to a small region (unless there is noise). If noise is there, it will be amplified. To avoid this, contrast limiting is applied. If any histogram bin is above the specified contrast limit, those pixels are clipped and distributed uniformly to other bins before applying histogram equalization. After equalization, to remove artifacts in tile borders, bilinear interpolation is applied.

For better result CLAHE and Morphology algorithm can be applied in combination. Morphological is efficient in getting malignant breast cancer region by reducing unwanted noise exist in the image. Both the algorithms combined has been able to give accuracy over 90%.

2.3 Segmentation:

Abdelali Elmoufidi, Khalid El Fahssi, Said Jai-Andaloussi and Abderrahim Sekkaki [3] have proposed "**Automatic Density Based Breast Segmentation for Mammograms by using Dynamic K-means Algorithm and Seed based region growing**" which provides brief overview in Segmentation Technique with the dynamic number of regions in mammogram.

The proposed method in this paper is to segment the image using dynamic K-means Algorithm and Seed based region growing techniques. Unlike other implementations like BI-RADs where the number of regions is static that is 5. In the proposed method the regions and number of regions are calculated dynamically.

By this method, more than one valuable ROI can be calculated. The K-means clustering algorithm is used to find the optimal number of centroids and the co-originates of those centroids. The Seed based Region growing is used for calculating the boundary of the ROI accurately.

The main goal of this study is to dynamically generate the number of regions in mammograms (so the number of regions in mammograms is variable from a mammogram image to another) and the input parameters values of region growing (seeds points and threshold's values) are automatically selected.

The proposed approach is,

1. Remove the digitization noise
2. Enhance the contrast of breast profile
3. Separate the breast profile from the background.

K-means clustering algorithm to generate the number of clusters and determine of seeds points and threshold's values for each cluster. Applied Seed Based Region Growing (SBRG) Techniques to classify pixels of the mammogram images into homogeneous sets plus detect the boundary of different breast tissue regions in mammograms.

As a result, an approach for the segmented and detected boundary of different breast tissue regions in mammograms, by using dynamic K-means clustering algorithm and Seed Based Region Growing (SBRG) techniques. The strong point of this study is that finding the number of regions, region selection and region growing can be automated. therefore, automatically segmented different breast tissue regions in mammograms with the accuracy of 92.87% of whole mammogram images.

Several researches have been done to develop CAD system to detect breast cancer. Tumour detection in mammograms includes denoising mammograms, followed by contrast enhancement, segmentation, feature extraction and classification.

2.4 Feature Extraction:

Ancy C A and Lekha S Nair proposed "**An Efficient CAD for Detection of Tumour in Mammograms using SVM**" which provides overview about an approach to develop a CAD system to detect breast cancer.

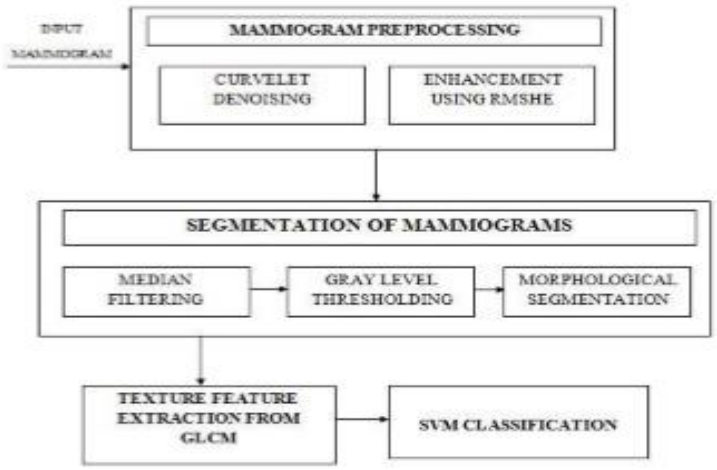

Fig 2.1: Block diagram of the Method.

The stages in proposed approach is shown above figure, where they apply Curvelet Denoising, RMSHE, Median Filtering, Gray Lebel Thresholding, Morphological Segmentation, GLCM for Texture extraction and Classifier for Classification.

Mammogram images usually contains the presence of salt and pepper and speckle noise. Denoising is to rectify this problem and can be done in spatial as well as in frequency domain. For denoising, Curvelet Denoising is used.

Contrast enhancement is the next step following denoising and is essential as mammograms have very low contrast and brightness preservation is also necessary. RMSHE (Recursive Mean Separate Histogram Equalization) Enhancement method is used for contrast enhancement.

Segmentation is done to find the ROI (Region of Interest) which contains masses and microcalcifications, and later on the suspicious region need to be separated from it. Static gray level thresholding and Morphological Segmentation techniques are used. Feature extraction is done to calculate the features that characterize the suspicious region and those features which are important for classification are selected and classified the feature into three types: geometric, texture and intensity features. Gray Level Coherence Matrix is used for extracting features.

The SVM classifier is used for Classification, the SVM (Support Vector Machine) classifier takes the extracted features from the image and then SVM provides soft-max'ed 3 element vector that provides the probability that the Mammogram belongs to which class. That is, Normal, Malignant and Benign.

As result, an accuracy of 96.6 % for using SEL weighted SVM, 91.5 % for using SVFNN and 82.1 % for using kernel SVM.

2.5 Classification:

M. A. Al-masni, M. A. Al-antari, J. M. Park, G. Gi, T. Y. Kim, P. Rivera, E. Valarezo, S.-M. Han, and T.-S. Kim have proposed "**Detection and Classification of the Breast Abnomaltites in Digital Mammograms via Regional Convolutional Neural Network**" that gives brief introduction to Regional CNN based classifier for Mammogram Classification.

Proposed approach is a novel computer-aided diagnose (CAD) system based on one of the regional deep learning techniques: a ROI-based Convolutional Neural Network (CNN) which is called You Only Look Once (YOLO) that contains four main stages: mammograms pre-processing, feature extraction utilizing multi convolutional deep layers, mass detection with confidence model, and finally mass classification using fully connected neural network (FC-NN). That is shown in figure 2.2.

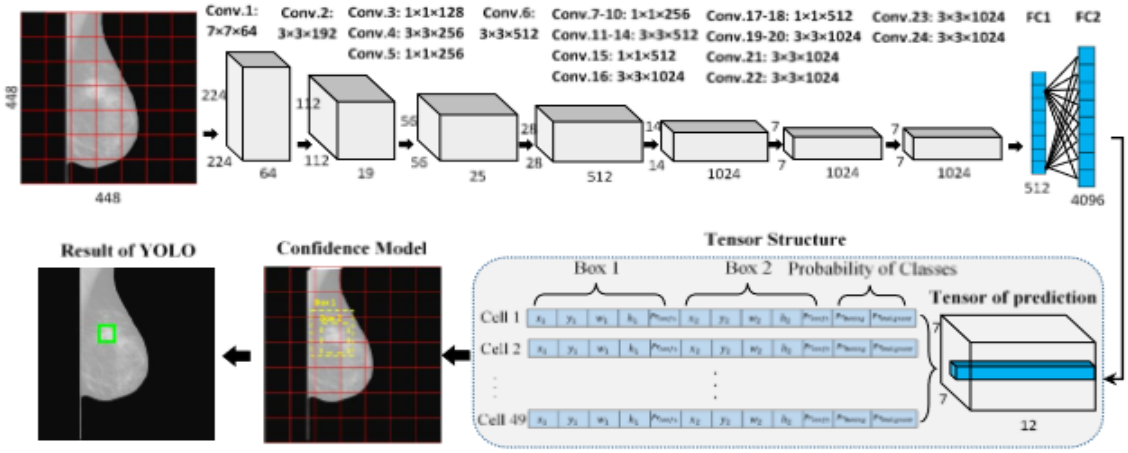

Fig 2.2: YOLO Model

In preparation of data, the multi-threshold peripheral equalization technique was applied to remove the effect of breast compression that occurred during the examining stage. This procedure improves the characteristics of the mammograms by eliminating the background and irrelevant data as presented in our previous work. The data are resized to a size of 448×448 and learned by our YOLO-based CAD.

YOLO is a unified system that is able to detect the potential ROIs and directly predict their class probabilities from an entire whole image. The proposed YOLO-based CAD highlights two main issues of finding out the mass locations and classification their types of benign or malignant. YOLO starts with dividing an input mammogram into N×N grid cells. Thus, each grid cell is responsible to detect the potential mass belonging to that cell. Two bounding boxes with their confidence scores are utilized to represent each grid cell. Confidence is expressed as the probability of the existing mass multiplied with the percentage of the intersections over union (IOU) between the ground truth box and the predicted one. Also, the detected mass is recognized as benign or malignant depending on the conditional class probability for the corresponding cell. Then, the confidence score for each specific class is estimated.

The proposed CNN uses 24 convolutional layers with a kernel size of 3×3, max-pooling layers with a size of 2×2, activation functions, and two fully-connected layers as shown in the above figure. Passed to the fully-connected network. Linear leaky rectified activation function is used for all layers and the Rectified Linear Unit (ReLU) is only used for the final layer. As result, 85.52 % overall accuracy is measured.

.Chapter 3

SYSTEM REQUIREMENT

3.1 Introduction

A Software Requirements Specification (SRS) is a description of a software system to be developed. It lays out functional and non-functional requirements, and may include a set of use cases that describe interactions between Users and the System. The software requirements specification document enlists enough and necessary requirements that are required for the project development. To derive the requirements, a clear and thorough understanding of the products to be developed or being developed is required. This is achieved and refined with detailed and continuous communications with the project team and customer till the completion of the software. Software applications to be used could be off-the-shelf applications modified to suit the project of they may be bespoke applications already available within the company. The overall purpose of the system specification documentation is to lay down exactly how the system is made up. Requirement analysis also called requirements engineering. It is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed. Requirement analysis is critical to the success of a system of software project. Conceptually, requirements analysis includes three types of activities.

● Eliciting Requirements for business process, documentation and stakeholder interviews. This is sometimes also called requirements gathering.

● Analysing Requirements for determining whether the stated requirements are clear, complete, consistent and unambiguous and resolving any apparent conflict.

● Recording Requirements for requirements to be documented in various forms, usually including a summary list and may include natural-language documents, use cases, or process specifications.

3.2 System Requirements

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as system requirements and are often used as a guideline as opposed to an absolute rule. Most Software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. A second meaning of the term of System requirements is a generalization of this first definition, giving the requirements to be met in the design of a system or subsystem. Typically, an organization starts with a set of Business requirements and then derives the System requirements from there.

There are two types of system requirements:

1. Software requirement.

2. Hardware requirement.

3.3 Software Requirements

A. Functional Requirements

● **Ubuntu 16.04.4 LTS:** Ubuntu is an open source operating system for computers. It is a Linux distribution based on the Debian architecture. It is usually run on personal computers, and is also popular on network servers, usually running the Ubuntu Server variant, with enterprise-class features. A default installations of Ubuntu contains a wide range of software. Many additional free required software packages are accessible from the built in Ubuntu Software Center as well as any other APT-based package management tools.

● **PyCharm:** PyCharm is an integrated Development Environment(IDE) used in computer programming, specifically for the python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, and integrated unit tester, integration with version control system, auto code completion, code generation, code folding, etc. PyCharm is cross-platform, with Windows,

macOS and Linux versions. The Community Edition is released under the Apache License, and there is also Professional Edition released under a proprietary license that has extra features.

● **OpenCV:** OpenCV is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage and is now maintained by Itseez. The library is cross-platform and free for use under the open-source BSD license.

● **Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides and object-oriented API for embedding plots into application using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. Matplotlib was originally written by John D. Hunter, as an active development community, and is distributed under a BSD-style license.

● **Numpy:** NumPy is a library for the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large collections of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. NumPy is open-source software and has many contributions.

● **Git:** Git is a version control system for tracking changes in computer files and coordinating work on these files among multiple people. It is primarily used for source code management in software development, but it can be used to keep track of changes in any set of files. As a distributed revision control system it is aimed at speed, data integrity, and support for distributed non-linear workflows. Git was created by Linus Torvalds in 2005 for the development of the Linux kernel, with other kernel developers contributing to its initial development.

B. Non Functional Requirements

● **Usability:** The system is made user interactive by providing a GUI (graphical user interface) using components like buttons, dropdowns, menus that are added to the window system to assist users.

● **Scalability:** Systems overall speed can be extended by integrating necessary modules to support GPU (Graphic Processing Unit) ‘s Cores for computation and to exploit the parallel processing ability of a system.

● **Portability:** Our application makes use of Python Programming Language that is interpreted using a Python Interpreter such as Cpython, Ipython etc So any Python Program can be executed in different platforms if the platform has Python Interpreter pre-installed. This makes our project compatible with many different computing platforms.

● Consistency and Extensibility: The codes can be reused and extended to higher accuracy and to extend the overall precision of the System.

3.4 Hardware Requirements

1. Processor: Intel Core i3 and above, 2 GHz

2. RAM: 4GB

3. Monitor: 1024*768 display resolution

4. Hard Disk Space: 100GB

3.5 Summary

Python offers higher cross functionality and portability as programs written in one platform can run across all other platforms given that there is a Python interpreter support. It’s easy to write and understand code in Python. The OpenCV (Open Computer Vision) has inbuilt image format supports such as PGM, JPEG, PNG and others. Numpy is well optimized Python higher dimensional matrix manipulation library. The matplotlib is a plotting library natively available in python that has higher level API’s to plot the graphs in real time.

Chapter 4

4.1 INTRODUCTION

The Mammogram is the most commonly employed technique for detection of breast cancer at an early stage. Previously steps to detect the Abnormality in Mammograms are, first is pre-processing of mammograms, second is Region of Interest (ROI) detection, third is feature extraction from ROI, Finally Classification that classifies the Tumors in Mammograms in as Benign or Malignant. The Cancer Diagnosis at the early stage is critical, Early detection increases the survival rate, increases treatment options and helps to improve the survivability of the patient. Radiologist's performance will be significantly improved in breast cancer detection when assisted by Computer Aided Detection (CADe) Systems. CAD can increase the diagnostic accuracy and efficiency with higher reproducibility. It has shown that the performance of a radiologist can be increased 5-45% by providing the radiologist with results from a CAD system as a second opinion. A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination. A Data Flow Diagram (DFD) is the graphical representation of the flow of data in an information system, modelling, its process aspects, A DFD is often used as a preliminary step to create an overview of the system without going into details which can be later elaborated. A data flow diagram illustrates how data is processed by a system in terms of inputs and outputs. As its name indicates its focus is on the flow of information, where data comes from, where it goes and how it gets stored. They provide a graphical representation of a system at any level of detail, creating an easy-to- understand picture of what the system does. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. UML sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic and are commonly used for analysis and design purpose. Sequence diagrams are time ordered interaction visual representation.

4.2 PROPOSED SYSTEM

The proposed system has 5 stages, as shown in figure 4.1,

1. Pre-processing

2. Segmentation

3. Feature Extraction

4. Classification

5. Stage Identification

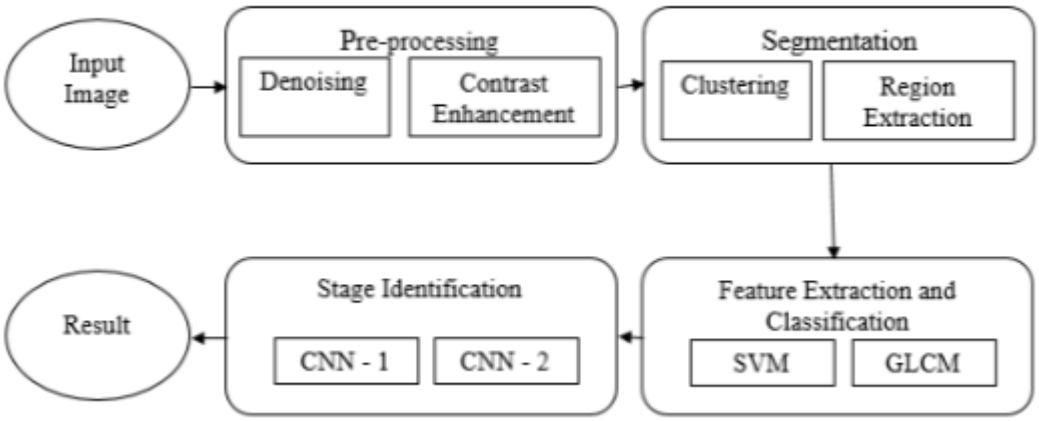

Fig 4.1: Block diagram of the proposed system

1. Pre-processing

Digital Mammogram images may contain some noise in the different regions that may affect Segmentation and/or Classification with intern may affect the accuracy, The mammogram will be denoised. The mammogram images are naturally having less contrast in the low light regions, these regions may also contain useful features that can be used to detect the tumors and micro calcifications. So the mammogram image will be enhanced using a Contrast Enhancement technique.

2. Segmentation

In this stage, the Region of Interest (ROI) will be calculated,the mammograms will have a black region on the left and right end sides. So these regions will be remove dusing threshold based morphology technique. Next the cropped image will be clustered to find the ROI that has micro calcifications and tumors.

3. Feature Extraction

In this stage, the meaning full features will be calculated from the Region of Interest (ROI). Feature extraction is done to calculate the features that characterize the suspicious region and those features which are important for classification are selected. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

4. Classification

Classification is used to classify the mammograms as Normal, Benign or Malignant based on the ROI calculated at Segmentation stage, and feature of ROI calculated in Feature extraction stage. The Normal mammograms are mammograms that doesn't have any malicious calcifications or indications of cancer. The Benign is non-spreading tumor otherwise a non-cancerous tumor whereas Malignant is a cancerous Mammograms that has malicious tumor indicating a spread.

5. Stage Identification

In this stage, the Stage of the Mammogram will be calculated using the Region of Interest (ROI). The Stages of cancer are ranges from 0 to 4 as shown below,

- Stage 0: Signs of cancer cells. (Earlier Stage)
- Stage 1: The cancer cells are contained to very limited area. (Earlier Stage)
- Stage 2: Cancer has begun to grow and spread. (Advance Stage)
- Stage 3: Cancer is invading tissue around the breast. (Advance Stage)
- Stage 4: Cancer has spread beyond the breast to other body.

The Stages are identified based on the features such as Texture of ROI, Shape and Spread of Abnormality. The first two stages will be in Benign stage and other 3 Stages are malignant as shown in figure 4.2 below.

Fig 4.2: Benign and Malignant Stages in Mammograms

4.3 DATA FLOW DIAGRAMS

A Data Flow Diagram (DFD) is a graphical representation of flow of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system. The DFD of the internal subsystem interaction is shown in the figure 4.3 below.

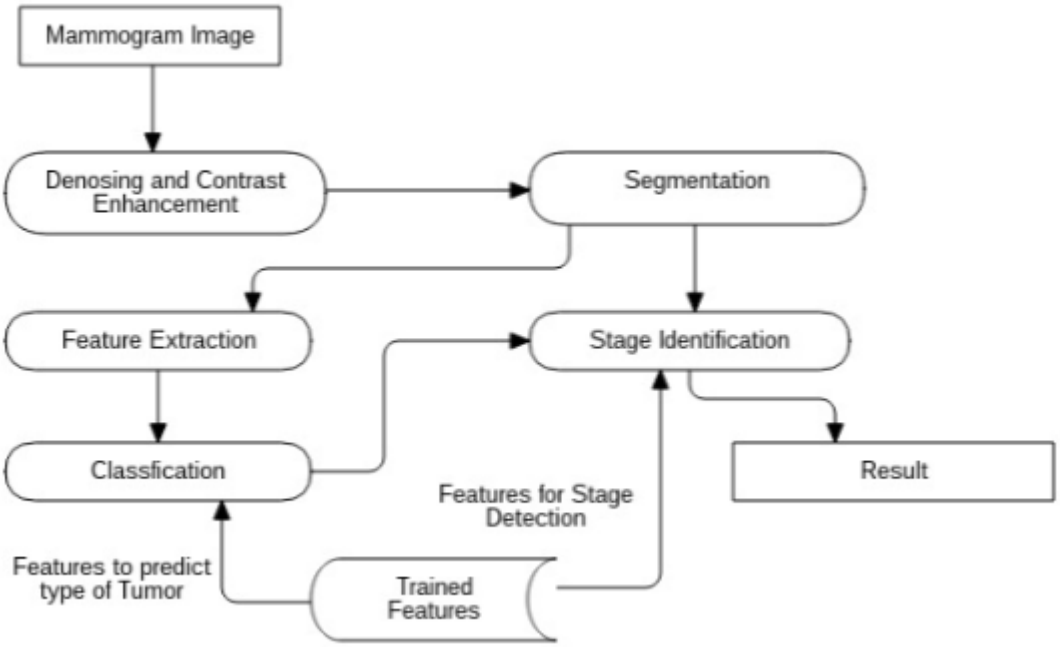

Fig 4.3: DFD for Internal Subsystem interaction

The figure 4.3 shows the interaction between the Subsystem in the proposed system, The Mammogram is first Denoised and Contrast Enhanced using Curvelet Transform

for denoising and Contrast - Limited Adaptive Histogram Equalization (CLAHE) for Contrast Enhancement.

The Enhanced Mammogram is Segmented to retrieve Region of Interest (ROI) from the Image. Mammogram contains a Black Region on both the right and left side, that black region will be removed using adaptive Thresholding algorithm and morphological technique. The cropped image will be clustered using a Adaptive K – Means algorithm, Using cluster centroids, The Seed based region growing algorithm will calculate the Region of Interest approximately.

The Region of Interest ‘s feature will be extracted using Grey-Level Co-Occurrence Matrix Algorithm. Those features are collectively called as Feature Vector. The Feature Vector is used by a SVM model that will Classify the Mammograms based on the Feature Vector of that Mammogram. The Output label of SVM model are Normal, Benign and Malignant. The Stages of Mammograms will be identified based on the Output Label from the SVM model and the ROI. Based on the Output label from SVM, The ROI will be passed to any one of Convolutional Neural Network Model that detects the Stage of Cancer in the Mammogram as shown in the Figure 4.4 below.

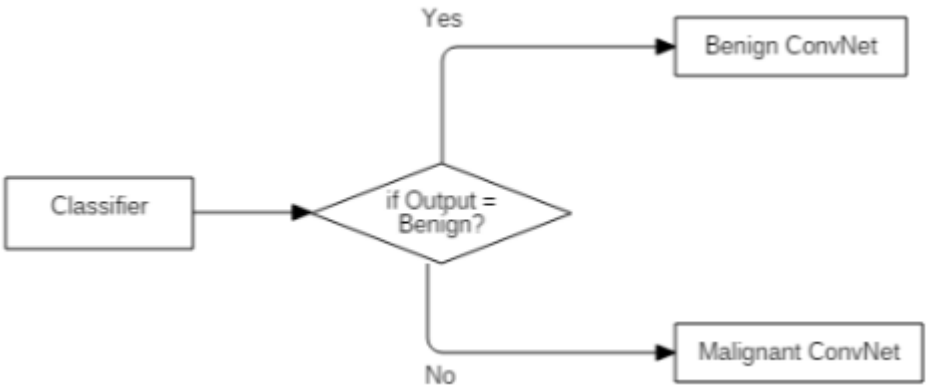

Fig 4.4: Control Flow in the Stage Identification

4.3.1 Pre-Processing

Level 0:

Fig 4.5: Level 0 DFD for Pre-processing

The Figure 4.5 shows the abstract input and output of the Pre-processing module, In the pre- processing module, the input is the digital mammogram image that may have noise and low contrast. The output of the pre-processing module is the noise free mammogram image with highlighted contrast. The contrast enhancement will be applied because the Mammogram images will contain meaningful features in the dark region of the image as well. First the Mammogram will be denoised (noise removal) and the Contrast enhancement will be applied and the resulting Image will be returned as the output.

Level 1:

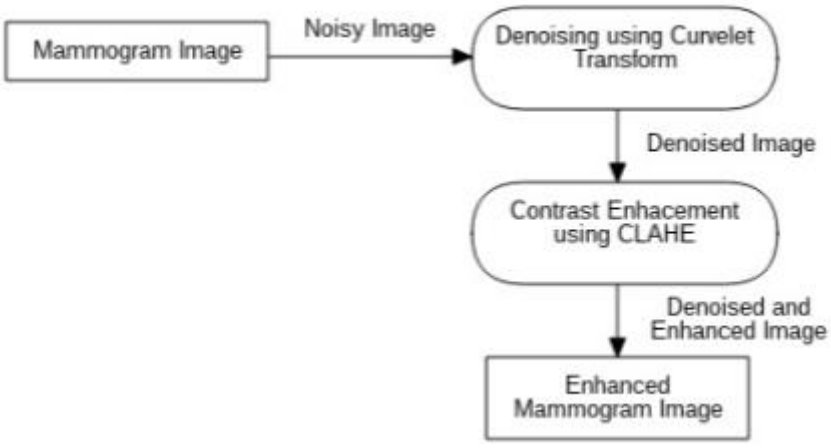

Fig 4.6: Level 1 DFD for Pre-processing

The figure 4.6 shows the subsystems in the Pre-Processing module and the interactions among those subsystems. As stated in Figure 4.5, The Mammogram is first denoised and contrast enhanced in that order. For Denoising the Curvelet

Transformation technique is used and Contrast-Limited Histogram Equalization (CLAHE) is used for Contrast Enhancement.

4.3.2 Segmentation

Level 0:

Fig 4.7: Level 0 for Segmentation

The Figure 4.7 shows the abstract input and output of the Segmentation module. Segmentation is performed to retrieve the partitioned portion of the mammogram image that has significant micro calcification or tumors. The input for the segmentation module is the Denoised and Enhanced Image from the Pre-Processing module, the output of Segmentation is the Region of Interest (ROI) of the mammogram. The Region of Interest (ROI) will have the decisive part of the Image. Segmentation is performed to remove the unwanted artifacts such as Track labels, Mammogram number or ambiguous regions that may affect the overall accuracy of the Proposed System.

Level 1:

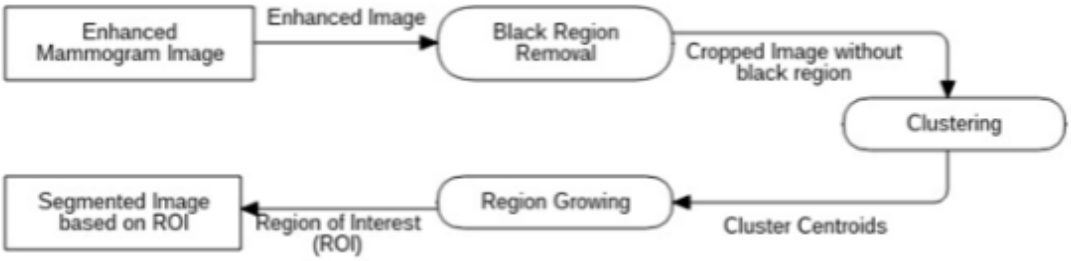

Fig 4.8: Level 1 DFD for Segmentation

The figure 4.8 depicts the internal subsystem of the Segmentation, the black region on both the right and left side of the Mammogram image is removed in Black Region

Removal. The cropped image is clustered to find the cluster centroids, one of the cluster centroid is part of ROI, then the Region Growing is applied to one of those cluster centroid to find the ROI of the Mammogram.

Level 2:

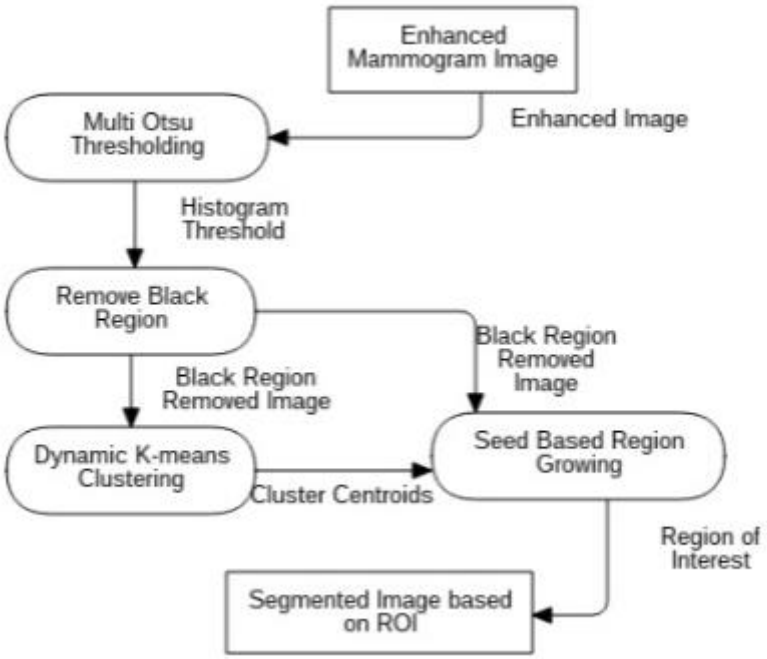

Fig 4.9: Level 2 DFD for Segmentation

The figure 4.9 shows the internal subsystems, and the data transformations between those systems and the Algorithms, Multi Otsu is an adaptive thresholding mechanism that produces one or more than one thresholds, the low intensity threshold is taken as pivotal point for Black region removal. The Dynamic K-Means is applied to cluster the regions in Mammograms to detect the ROI's centroid, that centroid becomes seed to detect the edges of ROI accurately. And then the ROI is cropped from the Mammogram Image.

4.3.3 Feature Extraction and Classification

Level 0:

Fig 4.10: Level 0 DFD for Feature Extraction and Classification

The figure 4.10 shows the abstract input and output of the Feature Extraction and Classification Module. The features of the cropped ROI from Segmentation is calculated in this module forming a Feature Vector. This Feature Vector is feed as input to a Classifier to classify the image as Normal, Benign or as Malignant.

Level 1:

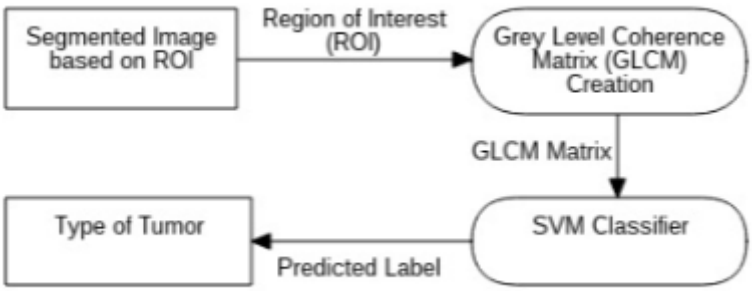

Fig 4.11: Level 1 DFD for Feature Extraction and Classification

The figure 4.11 depicts internal subsystems of the Feature Extraction and Classification, the Grey-Level Co-Occurrence Matrix (GLCM) is used to extract the features from the cropped ROI from the Segmentation Module, thus forming feature vector. The feature vector is used to Classify the Mammograms. The Classifier used in this module is SVM (Support Vector Machine). SVM feeds the feature vector as input and outputs the type of tumor or micro calcification.

4.3.4 Stage Identification

Level 0:

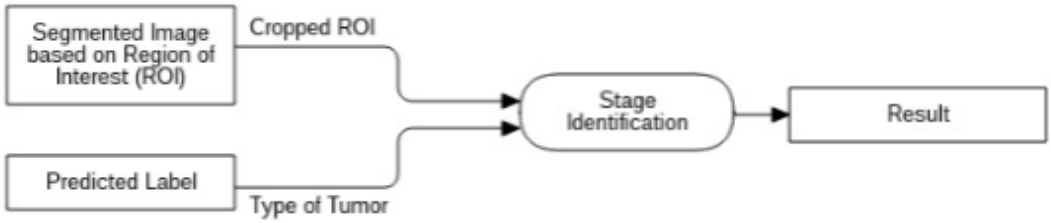

Fig 4.12: Level 0 DFD for Stage Identification

The figure 4.12 shows the abstract inputs and output of the Stage Identification. The Stage Identification accepts the ROI (Region of Interest) and The Predicted Type of Tumor from the Classification as input and outputs the Stage of Cancer. The output is completely depending upon the ROI's texture, curves and size. Stage Identification is critical because inaccurate Stage prediction can reduce the survivability of the patient.

Level 1:

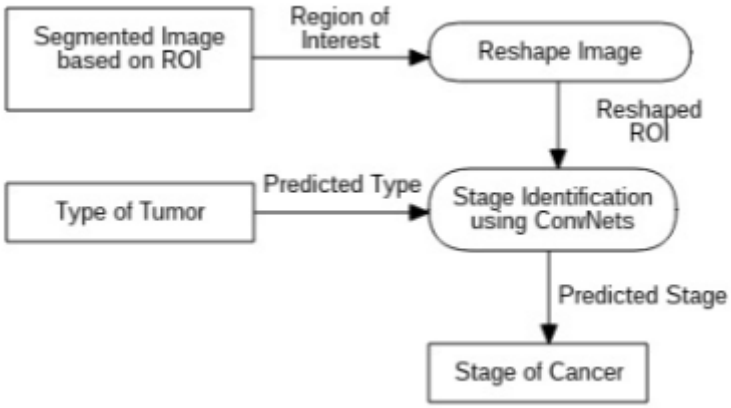

Fig 4.13: Level 1 DFD for Stage Identification

The figure 4.13 depicts the internal structure of the subsystems, and the data flow between those systems. The Segmented ROI from Segmentation is reshaped into a default fixed width and height. Based on the Type of Tumor predicted at the Classification Stage, any one of two Convolutional Neural Network will be selected.

The reshaped ROI will be fed as input to the selected Convolutional Neural Network (CNN). The selected CNN will output the stage of cancer. There are one Convolutional Neural Networks for each Type of Tumor.

4.4 Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. In software engineering, a system sequence diagram (SSD) is a sequence diagram that shows, for a particular scenario of a use case, the events that external actors generate, their order, and possible inter-system events. The sequence diagram is a suitable diagram to use to document a system's requirements and to flush out a system's design. The reason the sequence diagram is so useful is because it shows the interaction logic between the objects in the system in the time order that the interactions take place.

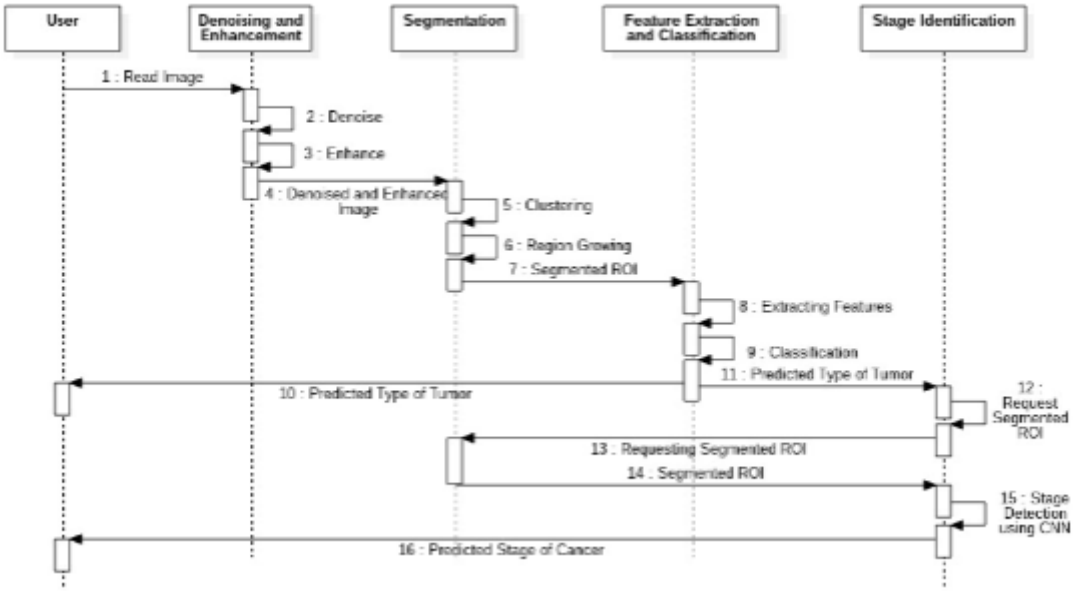

Fig 4.14: Sequence Diagram for the proposed system

The figure 4.14 shows one external user named “User” and 4 Subsystems in the proposed system, they are “Denoising and Enhancement”, “Segmentation”, “Feature Extraction and Classification” and “Stage Detection”. The first three subsystems communicate with each other in a straight manner, they perform their operation and they send the data generated to the neighbours. The denoising and enhancement will

be performed in 2 and 3 interactions in the first subsystem, Segmentation performs Clustering and Region Growing, The resulted ROI will have used to extract feature vector and then classified, predicted type of tumor will be sent to user. The stage detection accepts the predicted type of tumor and the ROI will be requested from the Segmentation subsystem, then the Stage detection will be performed using ConvNets . The predicted stage will be notified to user.

4.5 Summary

The existing system only classifies the tumors into Benign or Malignant. Whereas the proposed system classifies the tumors into Benign or Malignant and identifies the stage of the cancer. The radiologist can use the proposed Computer Aided Detection (CADe) System to improve their performance and to reduce the calamities caused by miss prediction. The mentioned DFD diagrams isolates the functions of the System into Subsystems which can be modularised effectively. Sequence diagram shows the interactions between those subsystems ordered by time, resulting in the series of steps that needs to be performed in the Proposed system.