VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi-590018, Karnataka, INDIA

PROJECT REPORT
on

"Feature Extraction and Classification of microcalcification clusters"

Submitted in partial fulfillment of the requirements for the VII Semester

Bachelor of Engineering
IN
COMPUTER SCIENCE AND ENGINEERING

For the Academic year
2018-2019
BY

HARSHITHA C.	**1PE15CS059**
LOKESH SHANMUGA	**1PE15CS075**
MEGHASHYAM	**1PE15CS084**

Under the Guidance of
Prof. Sangeetha R
Assistant Professor, Dept. of CSE
PESIT-BSC, Bengaluru-560100

PES INSTITUTIONS

Department of Computer Science and Engineering
PESIT BANGALORE SOUTH CAMPUS
Hosur Road, Bengaluru -560100

PESIT BANGALORE SOUTH CAMPUS

Hosur Road, Bangalore -560100

Department of Computer Science and Engineering

CERTIFICATE

Certified that the project work entitled ***"Feature Extraction and Classification of Microcalcification Clusters"*** *is a bonafide work carried out by* ***Harshitha C, Lokesh Shanmuga and Meghashyam*** *bearing USN* ***1PE15CS059, 1PE15CS075 and 1PE15CS084*** *respectively, students of* ***PESIT Bangalore South Campus*** *in partial fulfillment for the award of* ***Bachelor of Engineering*** *in* ***Computer Science and Engineering*** *of the* ***Visvesvaraya Technological University****, Belagavi during the year 2018-2019. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated and the project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.*

Signatures:

Project Guide **Prof. Sangeetha R** Assistant Professor, Dept. of CSE PESIT-BSC, Bengaluru	Head Dept of CSE **Dr. Sandesh B J** Professor, Dept. of CSE, PESIT-BSC, Bengaluru	Director/Principal **Dr. J. Suryaprasad** PESIT-BSC, Bengaluru

External Viva

Name of the Examiners	**Signature with date**
1. _____	_____
2. _____	_____

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned my effort with success.

We are indebted to our Guide, **Prof. Sangeetha R**, Assistant Professor, Department of Computer Science and Engineering, PESIT - Bangalore South Campus, who has not only coordinated our work but also given suggestions from time to time.

We are also extremely grateful to our Project Co-ordinators, **Prof. Keerti Torvi,** Assistant Professor**, Prof. Sangeetha R,** Assistant Professor**, Prof. Jyoti Desai,** Assistant Professors, Department of Computer Science and Engineering, PESIT Bangalore South Campus, for their constant support and advice throughout the course of preparation of this document.

We are greatly thankful to **Dr. Sandesh B J**, Professor and HOD, Department of Computer Science and Engineering, PESIT Bangalore South Campus, for his able guidance, regular source of encouragement and assistance throughout this project.

We would like to express our immense gratitude to **Dr. J. Suryaprasad,** Director and Principal, PESIT Bangalore South Campus, for providing us with excellent infrastructure to complete our project work.

We gratefully acknowledge the help lent out to us by all faculty members of the Department of Computer Science and Engineering, PESIT Bangalore South Campus, at all difficult times. We would also take this opportunity to thank our college management for the facilities provided during the course of the project. Furthermore, we acknowledge the support and feedback of my parents and friends.

Harshitha C (1PE15CS059)
Lokesh Shanmuga (1PE15CS075)
Meghashyam (1PE15CS084)

ABSTRACT

Breast cancer is the most common invasive cancer in women and the second main cause of cancer death in women after lung cancer. With breast cancer, early detection is key. The earlier the disease is diagnosed , the better the outcome with treatment.

This project aims in detecting the breast cancer at a very early microcalcification stage itself rather than at tumor stage using different efficient techniques of extraction of the features such as brightness, contrast, size, shape from microcalcification clusters and then classifying the microcalcification clusters into either benign or malignant cancer cells with the help of the extracted features.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

1.1 Purpose

- Breast cancer is the most common invasive cancer in women. With breast cancer, early detection is key. The earlier the disease is diagnosed , the better the outcome with treatment.This project aims in detecting the breast cancer at a very early micro calcification stage itself rather than at tumor stage.
- The radiological definition of cluster microcalcification is the presence of more than three microcalcification in 1 cm2 area.A given cluster of microcalcification be associated with malignant and benign case.Distinguishing between malignant and benign cluster is a difficult and time consuming task for radiologists.Only 20 percent to 30 percent biopsy cases recommended by the radiologists turn out to be malignant.
- Its of crucial importance to design the classification method in such a way to obtain high level of TPF(True Positive Fraction) while maintaining the FPF (false -positive fraction)at its minimum.It has been shown that computerized detection and classification thus outperform radiologists detection and classification

1.2 Scope

- This main objective is to detect the breast cancer at a very early microcalcification stage.
- Many features such as brightness, contrast, size, shape, and texture are extracted from microcalcification clusters using different efficient feature extraction techniques.
- And then classifying clusters into either benign or malignant cancer cells with the help of the extracted features.

1.3 Literature Survey

- In the Research paper- Classification of Benign and Malignant Breast Masses Based on Shape and Texture Features in Sonography Images by Fahimeh Sadat Zakeri and Hamid Behnam and Nasrin Ahmadinejad various shape features were extracted.

1.Eccentricity-feature: It is a scalar that describes the eccentricity of the ellipse which has the same second moment as the mass region. The eccentricity value is obtained from the ratio of the distance between the foci of this ellipse to its major axis length. 2.Solidity-feature: It is a scalar that describes the proportion of the pixels in the mass region to the pixels in the convex hull including the mass. It is computed as (Masss area)/ (Convex hulls area). 3. DifferenceArea-Hull-RectangularX: It is absolute value of convex hulls area minus convex RectangularXs area. 4. DifferenceArea-Mass-RectangularX: It is absolute value of convex RectangularXs area minus masss area. 5. Cross-correlation-left: It is the cross correlation value between the RectangularX region and the Left-side region. 6. Cross-correlation-right: It is the cross correlation value between the RectangularX region and the Right-side region.

- Surround Region Difference Method proposed in Reasearch paper by Jong Kook Kim and Hyun Wook Park is summarized as follows.

  The surrounding region-dependence matrix contains the texture information of an ROI. The texture coarseness or fineness of an image can be interpreted as the distribution of the elements in the matrix. If a texture is smooth the distribution of elements should be concentrated on or near upper left corner of the matrix. If a texture has fine details the distribution of elements should be spread out along the diagonal of the matrix. The distribution of elements tends to spread to the right and/or lower right corner of the matrix for positive ROI's containing clustered microcalcifications.

- Gray Scale level features:

  The texture features are extracted using GLCM(grey level co occurrence matrix). The GLCM is a two dimensional array which takes into account the specific position of a pixel relative to other pixels The GLCM is a tabulation of how often different combination of pixel brightness values occur in an image. The texture descriptors derived from GLCM are cluster shade, contrast, energy and sum of square variance.

- Comaprision of Classifiers:(Competitive Study of Classification Techniques on Breast Cancer FNA Biopsy Data - Daniele Soria ,Jonathan M. Garibaldi,Elia Biganzoli,Ian O. Elli)

1 Decision tree algorithms: Frequently used decision tree algorithms are ID3,C4.5,and CART.CART (Classification and Regression Trees)provides better accuracy in classifying the breast cancer data sets than ID3,C4.5 algorithms. C4.5 Classifier : C4.5 is an algorithm used to generate a decision tree and these trees can be used for classification. C4.5 uses the concept of information gain to make a tree of classifactory decisions with respect to a previously chosen target classification. The output of the system is available as a symbolic rule base. The cases are scrutinized

for patterns that allow the classes to be discriminated. These patterns are then expressed as models, in the form of decision trees, which can be used to classify new cases.Since for the real world databases the decision trees become huge and di□cult to understand and interpret

2 Multilayer Perceptron Classifier: A multilayer perceptron is a forward artificial neural network model that maps sets of input data onto a set off appropriate output. It has distinctive characteristics: 1. Model of neuron in the network includes a nonlinear activation function. 2. The network consists one or more layers of hidden neurons that are not part of input and output of the network. Error back-propagation algorithm is used in this type of classifier

3 Nave Bayes Classifier: A nave bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions It is based on two simplifying common assumptions: firstly, it assumes that the predictive attributes are conditionally independent given the class and secondly, the values of numeric attributes are normally distributed within each class. Bayes theorem is given as : p(C=c/X=x)= p(C=c)p(X=x/C=c)/p(X=x)

4 Support Vector Machine : SVM is a technique based on the statistical learning theory.It seperates two classes by determining the linear classifier that maximizes the margin . According to the paper on comparitive study of classification by Haowen You and Gearge Rumbe, accuracy of this classifier remained constant at 62.74% and progressively better predictions above 90 percent were determined. Initially in this method SVM was trained by varying gamma values and range was selected and classifier was evaluated including functions like polynomial, sigmoid and radial bias function

1.4 Existing Systems

Research on breast cancer using digital image processing is not new but lack of proper methods for early detection at microcalcification stage is still a challenge to medical domain.Most of the research work done till now detects the breast cancer at tumor stage and are not accurate at 100% and leads to false positive or false negative results which are highly dangerous.

1.5 Proposed System

The proposed model uses a combination of different highly efficient techniques of digital imaging to extract different features from microcalcification clusters and then classifying clusters into either

benign or malignant cells.

The proposed system will be of high accuracy leading to true positive and true negative results.

1.6 Statement of the problem

Feature Extraction and Classification of micro-calcification clusters. The purpose of feature extraction is to reduce the data set by measuring some properties, or features that distinguish one input pattern from the other. The extracted features provide characteristics of the input type to the classifier considering the relevant description of the image.

1.7 Summary

This chapter gave a brief introduction on what exactly the proposed system is. It also covered the future scope and the demand of this product.It shows the main features overcoming the drawbacks of the existing system.

Chapter 2

System Requirements Specifications

2.1 Software Requirements Specifications

Requirement specification is the movement of interpreting the data assembled amid investigation into prerequisite report.
Software requirements specifications are the detailed enlisting of all necessary requirements that arise in the project. The aim of having these requirements is to gain an idea of how the project is to be implemented and what is to be expected as a result of the project. The sections in this chapter deal with the various kinds of software, hardware and other functional and non functional requirements of the project. A brief description of the various users of the system is also mentioned.

2.1.1 Operating Environment

This section gives a brief about the hardware and software prerequisites for the project.

Hardware Requirements

- **Processor**: 2GHz or faster processor
- **RAM**: 2 GB(64bit)
- **Storage**: 250GB of available hard disk space
- Other general hardwares such as a mouse and keyboard for inputs and a monitor for display.

Software Requirements

- **Operating system**: Ubuntu 14.04 and above or Windows 7 and above
- **Programming languages**: Python
- **Documentation**: Overleaf

2.1.2 Functional Requirements

Functional requirements are a formal way of expressing the expected services of a project. We have identified the functional requirements for our project as follows:

- The system should be able to gather data from datasets.
- The system should be able to check for correctness of data extracted.
- The system should be able to efficiently extract features from the datasets.
- The system should be able to have the capacity to decide the contribution of each attribute towards the decision made by the predictor.
- The system should be able to classify the examples accurately.

2.1.3 Non-Functional Requirements

Non functional requirements are the various capabilities offered by the system. These have nothing to do with the expected results, but focus on how well the results are achieved.

- Reliability :The subsystem gives results with an high accuracy score. The systems are highly accurate subject to the proper working of each component involved.
- Scalability: The system should be able to run over huge datasets as well as small datasets and should be able to produce accurate results in both the cases
- Performance : The system will yield quick results for new queries.
- Portability : This is required when the computer, which is facilitating the framework stalls out because of few issues, which requires the system to be taken to another computer.
- Re-usability : The degree to which existing applications can be reused in new application. The predicted output could be reused in many fields.

2.1.4 User Characteristics

There is only one type of user associated with the system:

- Medical Personnel: He/She may use this system to detect the disease at the early stage which can be useful for providing proper treatment to the patients.

2.1.5 Applications

Medical domains: This system is used in medical domain for the diagnosis of breast cancer at the very early stage and will be useful in providing the proper treatment.

2.1.6 Summary

This chapter discussed the basic software and hardware requirements. More importantly it discusses the functional and non-functional requirements.

Chapter 3

High Level Design

This section mainly covers the design technique of the entire system.:

- The Feature extraction phase and classifier.

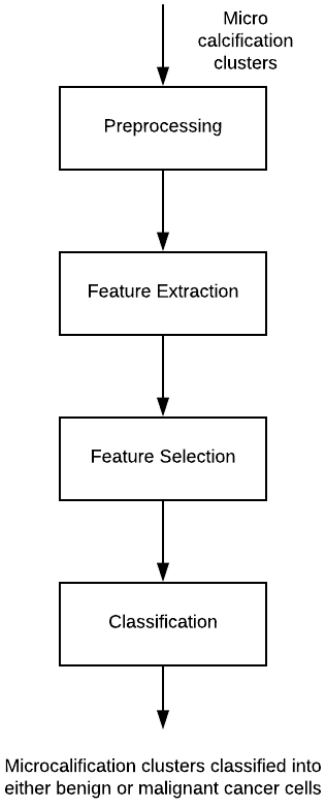

Figure 3.1: High Level Design

3.1 Design Approach

Here are two methodologies for software designing:

- Top-down Design:It takes the entire programming framework as one entity and after that disintegrates it to accomplish in excess of one subsystem or some components based on few attributes.
- Bottom-up Design: The model begins with most particular and essential components. It accedes with making more elevated amount out of subsystems by utilizing essential or lower level

components.

As mentioned above the project requires two main modules to be implemented. Each module has its own components to be developed. We use bottom-up design strategy in this product design phase as we start designing the basic components in each module and finally we interlink both the modules to get the final product.

3.2 System Architecture

System architecture is the conceptual model that define the structure, behavior and views of a system. It is a formal description and representation of a system, the basic structure of proposed system is depicted in the below figure

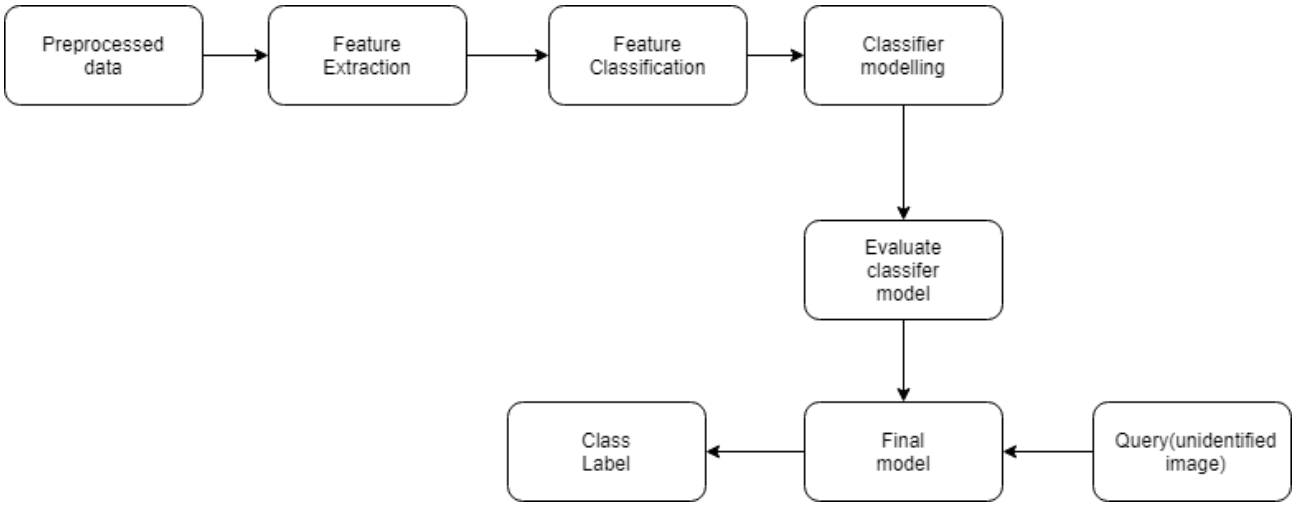

Figure 3.2: System Architecture

3.3 Data Flow Diagram

Data Flow Diagram is the starting point of the design phase that functionally decomposes the requirement specification. A DFD consists of a series of bubbles joined by lines. The bubbles represent data transformation and the lines represent data flows in the system. A DFD describes what data flow rather than how they are processed, so it does not include hardware, software and data structure.

A Data Flow Diagram is a graphical representation of the "flow" of data through an information system. DFD's can also be used for the visualization of data processing (Structured Design). A data flow design is a significant modeling technique for analyzing and constructing information processes.A DFD can be referred to as a Process Model.

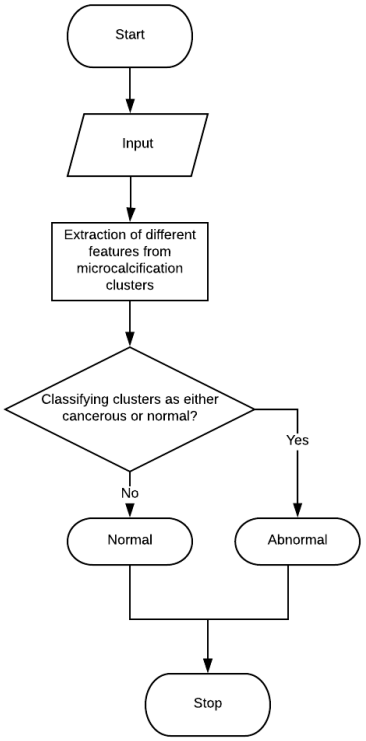

Figure 3.3: Data Flow Diagram

3.4 Summary

In this chapter we discussed the different design patterns which can be used in any product development cycle. We also discussed the flowchart which shows the data flow between various components. This chapter even described the high level design.

Chapter 4

Detailed Design

4.1 Purpose

The purpose of the detailed design is to plan our system to meet the requirements specified at the start. In the detailed design we see what is the input data for each model, how the model implementation is carried out and how the output is interpreted.

4.2 Module 1: Feature Extraction

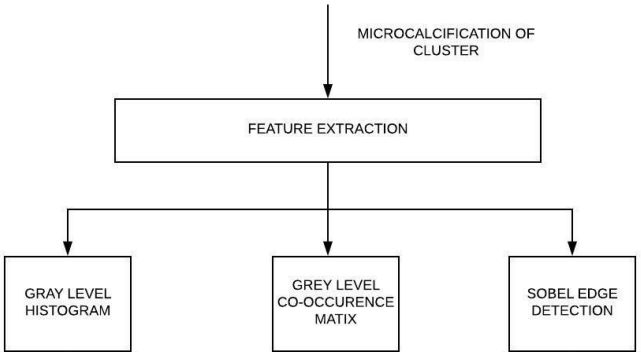

Figure 4.1: Detailed Design

The dataset we used is the popular MIAS database. After the preprocessing phase the images are input to the feature extraction module.The feature extraction module uses 3 methods to extract meaningful features about the image.These methods are
1.Obtaining statistical moments of the gray level histogram of the image.
2.Extracting entropy,uniformity and other features from the Gray Level Co-occurrence matrix constructed from the image.
3.Apply sobel edge detection mask.
Once the features are extracted,if required feature selection process is carried out.

4.3 Module 2: Classification

Classification is the process of grouping of entities with the similar attribute under one class label. This is necessary because it validates our hypothesis on the efficiency of our model. Classification in

our case is a binary classification whether the microcalcification cluster is either benign or malignant. We would use a Naive Bayes classifier to do the classification job.

4.4 Summary

The modules of the system are the basic units of functionality of a system.They interact with the user to complete the working of the system. The various modules that are implemented along with the expected inputs and outputs are mentioned in this chapter. Defining these parameters will help the user as well as the developer get a clear understanding of the project in addition to easing out the testing process.

TEAM INFORMATION

1. NAME: Harshitha C

USN: 1PE15CS059

CONTACT NUMBER: 9008306686

EMAIL ID:harshithareddy097@gmail.com

ADDRESS:#2,Lakshminarayanapura, Huskur Post, Anekal Taluk, Bangalore 560099

2. NAME: Lokesh Shanmuga

USN: 1PE15CS075

CONTACT NUMBER: 9663140921

EMAIL ID: shanmuga.lokesh0@gmail.com

ADDRESS: #8,4th main, 12th cross M.C Layout , Vijaynagar, Bangalore-560040

3. NAME: Meghashyam

USN: 1PE15CS084

CONTACT NUMBER: 8892146794

EMAIL ID: kushalq@gmail.com

ADDRESS: #17/23 3rd main, Shakambri nagar, J.P Nagar 1st Stage, Bangalore