Multi-level Parallelization For Accurate And Fast Medical Image Retrieval

Keith Sasala Chikamai

209540589

A thesis submitted to
the University of KwaZulu-Natal,
College of Agriculture, Engineering and Science,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Supervisor: Prof Serestina Viriri
Co-Supervisor: Prof Jules-Raymond Tapamo

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

December 2016

I, Keith Sasala Chikamai , declare that:

(i) The research reported in this thesis, except where otherwise indicated, is my original research.

(ii) This thesis has not been submitted for any degree or examination at any other university.

(iii) This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced:

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

Candidate: Keith Sasala Chikamai

Signature: _____ Date: _____

As the candidate's supervisors we approve the submission of this thesis for examination.

Supervisor: Prof Serestina Viriri, Co-Supervisor: Prof Jules-Raymond Tapamo

Signature: _____ Signature: _____

Date: _____ Date: _____

Abstract

Breast cancer is the most prevalent form of cancer diagnosed in women. Mammograms offer the best option in detecting the disease early, which allows early treatment and by implication, a favorable prognosis. Content-based Medical Image Retrieval (CBMIR) technique is increasingly gaining research attention as a Computer Aided Diagnosis (CAD)) approach for breast cancer diagnosis. Such systems work by availing mammogram images that are pathologically similar to a given query example, which are used to support the diagnostic decision by referential basis. In most cases, the query is of the form "return k images similar to the specified query image". Similarity in the Content-based Image Retrieval (CBIR) context is based on the content of images, rather than text or keywords. The essence of CBIR systems is to enable indexing of pictorial content in databases and eliminating the drawbacks of manual annotation. CBMIR is a relatively young technology that is yet to gain widespread use. One major challenge for CBMIR systems is bridging the "semantic gap" in the description of image content. Semantic gap describes the discord in the notion of similarity between the descriptions of humans and CBMIR systems. Low accuracy concerns inhibit the full adoption of CBMIR systems into regular practice, with research focusing on improving the accuracy of CBMIR systems. Nonetheless, the area is still an open problem.

As a contribution towards improving the accuracy of CBMIR for mammogram images, this work proposes a novel feature modeling technique for CBMIR systems based on classifier scores and standard statistical calculations on the same. A set of gradient-based filters are first used to highlight possible calcification objects; an Entropy-based thresholding technique is then used to segment the calcifications from the background. Experimental results show that the proposed model achieves a 100% detection rate, which shows the effectiveness of combining the likelihood maps from various filters in detecting calcification objects.

Feature extraction considers established textural and geometric features, which are calculated from the detected calcification objects; these are then used to generate secondary features using the Support Vector Machine and Quadratic Discriminant Analysis classifier. The model is validated through a range of benchmarks, and is shown to perform competitively in comparison to similar works. Specifically, it scores 95%, 82%, 78%, and 98% on the accuracy, positive predictive value, sensitivity and specificity benchmarks respectively.

Parallel computing is applied to the task of feature extraction to show its viability in reducing the cost of extraction features. This research considers two technologies for implementation: distributed computing using the message passing interface (MPI) and multicore computing using OpenMP threads. Both technologies involve the division of tasks to facilitate sharing of the computational burden in order to reduce the overall time cost. Communication cost is one penalty

implied with parallel systems and a significant design target where efficiency of parallel models is concerned. This research focuses on mitigating the communication overhead for increasing the efficacy of parallel computation; it proposes an adaptive task assignment model dependent on network bandwidth for the parallel extraction of features. Experimental results report speedup values of between $4.7x$ and $10.4x$, and efficiency values of between 0.11 and 0.62. There is a positive increase in both the speedup and efficiency values with an increase in the database size. The proposed adaptive assignment of tasks positively impacts on the speedup and efficiency performance of the parallel model. All experiments are based on the mammographic image analysis society (MIAS) database, which is a publicly available database that has been widely used in related works.

The results achieved for both the mammogram pathology-based retrieval model as well as its computational efficiency met the objectives set for the research. In the domain of breast cancer applications, the models proposed in this work should positively contribute to the improvement of retrieval results of computer aided diagnosis/detection systems, where applicable. The improved accuracy will lead to higher acceptability of such systems by radiologists, which will enhance the quality of diagnosis both by reducing the decision-making time as well as improving the accuracy of the entire diagnostic process.

List of Publications

[1] Chikamai, K., Viriri S. and Tapamo, J-R. (2015), In: "The effectiveness of combining the likelihood maps of different filters in improving detection of calcification objects," *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pp. 30-36.

[2] Chikamai, K., Viriri S. and Tapamo, J-R. (Accepted), "Mammogram Content-based Image Retrieval Based on Malignancy Classification," *Intelligent Data Analysis*, 21(5), September, 2017.

[3] Chikamai, K., Viriri S. and Tapamo, J-R. (Submitted), "On the reduction of Computational response times of feature extraction through Optimal Application of Cluster and Multicore Computing," *International Journal of Computational Science and Engineering (IJCSE)*.

Acknowledgements

My utmost gratitude to God, the beginning and end of it all.

I am grateful for the motivational, moral and financial support of my parents, the moral support of my brothers: Kevin and John as well as the contribution, in all its nuances, of my extended family and friends.

I acknowledge and express my gratitude to the Centre for High Performance Computing (CHPC) South Africa for the provision of their resources, thereby enabling the development of the parallel processing model of this project.

I thank my supervisors for their all-rounded guidance that went beyond their stipulated responsibilities, without which this work would not have been completed and additionally, within the timeframe it was.

Contents

List of Figures

List of Abbreviations

ANN Artificial Neural Network

BI-RADS Breast Imaging Reporting and Data System

CORBA Common Object Request Broker Architecture

CBIR Content-based Image Retrieval

DICOM Digital Imaging and Communications in Medicine

DDSM Digital Database for Screening Mammography

ESI Electrical Impedance Spectroscopy

FIR Finite Impulse Response

FSFS Forward Selection Forward Search

GPU Graphical Processing Unit

IRMA Image Retrieval in Medical Applications

k**-NN** $k-$Nearest Neighbor

LDA Linear Discriminant Analysis

MPI Message Passing Interface

MIAS Mammographic Image Analysis Society

MIMD Multiple Instruction Multiple Data

PACS Picture Archive and Communication System

PCA Principal Component Analysis

PPV Positive Predictive Value

QDA Quadratic Discriminant Analysis

HIS Hospital Information System

ROI Region of Interest

SVM Support Vector Machine

SIMD Single Instruction Multiple Data

TP True Positive

Chapter 1

Introduction

Information forms an important and ubiquitous resource in any institution or domain. Virtually all systems, physical and logical, maintain an information flow of some kind. The need for structured and efficient management of information has seen either the manual file systems, the Database Management Systems (DBMS), or both, become an inherent part of virtually all organizations. Regardless of the specific implementation used, the information cycle generally involves four processes: input, processing, storage and retrieval. The efficacy of each of these subcomponents has a direct bearing on the overall performance of the entire system. The research domain is interested in optimizing each of these components. For instance, superior hardware devices are being developed to better capture and represent phenomena, processing devices are being improved to make them faster, storage devices are being enhanced to allow more capacity for storage and retrieval mechanisms are being improved to enable faster and accurate retrieval of stored data.

The traditional notion of information handling with regards to DBMS commonly made reference to textual data, even when the actual data was multimedia in nature. For instance, while pictorial content has been stored in databases, access to it has been through textual tags intuitively coined to reference the pictorial data. In radiology departments, screening procedures for patients culminate in X-ray images. These images traditionally were annotated using textual tags for future reference. However, this method presents drawbacks such as

- Ambiguity of expression - multiple and confusing textual descriptions might be possible for

the same image

- Incompleteness of information - the image data may not be exhaustively captured by the textual description
- Limited dynamism - updating information about a given image would be intrinsically strenuous and laborious.

CBIR is a relatively recent technique developed to address this shortcoming; it refers to the notion of retrieving images based on their pictorial content rather than textual annotations. In the medical domain, CBIR plays an important role of retrieving images of the same form, anatomical region or pathology. In the latter case, it is used as a diagnostic aid tool. This involves the retrieval of a set of similar images to a given query, enabling a radiologist to diagnose the case in hand by referencing the retrieved set of images which, ideally, carry a similar pathology. In spite of this, these systems are yet to gain widespread use because of accuracy concerns. Research effort is ongoing to improve the accuracy of CBIR algorithms, to enable them capture the relevant "medical characteristics" while minimizing irrelevant information. The capture and representation of information from images is known as feature extraction, and has significant bearing on the accuracy CBIR algorithms. While various levels of satisfaction has been achieved in improving feature extraction, there remains room for improvement. Furthermore, some CBIR algorithms are computationally demanding, necessitating the need for design options that will reduce their computational complexity. This research study will focus in developing an efficient performance model for mammogram image retrieval, in the sense of improving feature extraction accuracy as well as reducing computational complexity.

1.1 Motivation

The availability of novel effective features presents an opportunity at improving the characterization of mammogram images. One probable feature is the Local Directional Pattern (LDP), a novel feature that encodes textural information with regard to orientation [1]. LDP is an improvement to the Local Binary Pattern; it is more stable in the presence of noise and non-monotonic illumination variation, which are characteristic of many mammogram images. Another opportunity for improving the accuracy of feature extraction lies in combining existing features [2]. Authors have investigated the efficacy of various singular features for characterizing mammogram images to considerable satisfaction. While there have been efforts to combine multiple features for characterization, there still exist opportunities for more potentially efficacious combinations.

Furthermore, while a given feature set might fail to achieve acceptable classification performance, transformation of the same might yield a feature set that is more correlated with a certain class, thereby improving the accuracy [3].

The application of parallel processing techniques has not been extensively considered in mammogram retrieval algorithms, even when the algorithms proved to be computationally expensive. Parallel processing techniques have been proven to reduce the runtime of tasks, provided that these tasks are coarsely grained. The intrinsic nature of a number of image processing algorithms used for enhancement and feature extraction operations allows for their concurrent application, which can reduce the overall computational times. The use of grid-based systems has also been found to improve the timeliness of retrieval algorithms in practical applications [4, 5]. Perez et al. [6] deploy on a gLibrary/DRI grid platform their CAD system for diagnosis of 6 pathological lesions. Oliveira et al. [7] also employ the Grid network for retrieval tasks over a medical database comprising of Magnetic Resonance Images (MRI) of two anatomical regions: Sagittal knee and axial head. Positive results from these works and a few related others demonstrate the potential of parallel processing as a means of reducing computational complexity.

1.2 Problem statement

Towards arriving at diagnostic decisions, radiologists analyze mammogram images for the presence of pathological objects, which may be indicative of certain diseases such as cancer. The nature of the pathological objects, if they exist, might give out additional information such as severity and extent of the disease. In difficult-to-diagnose cases, radiologists might make reference to historical cases when processing a given case. Furthermore, radiologists usually consider previous cases when monitoring the trend of the disease for a particular patient. CBIR-based CADe/x systems provide a diagnostic aid to radiologists in this scenario, by retrieving cases that are "pathologically similar" to the case in hand.

The goal of medical information is defined as the need to "deliver the needed information at the right time, the right place to the right persons in order to improve the quality and efficiency of care processes" [8]. The failure to adopt CADe/x systems in widespread practical routines is premised on that goal and mostly attributed to their low accuracy rates. One concern about these algorithms is their failure to bridge the semantic gap. This means that the features used for characterization fail to accurately capture the pathological objects as desired by radiologists [8]. Researchers have attempted to explore various feature sets to improve the accuracy of characterization to various

levels of success. However, this research domain is still open, and there exist opportunities for improvement.

The retrieval problem is computationally expensive, where a large database is involved. Some algorithms are intrinsically computationally demanding. This can be compounded with further scalability of the image database, which is an inevitable trend. This poses a challenge since diagnostic decisions need to be made fast in order to allow early commencement of treatment if need be.

1.3 Thesis objectives

This objectives of this research are to,

1. Conduct a critical appraisal of existing approaches used for feature extraction of mammogram images
2. Develop a novel algorithm for characterization of mammogram features that will improve the accuracy of content-based retrieval of mammogram images.
3. Develop a parallel processing model that will ameliorate the time complexity of the feature extraction process

1.4 Reasearch questions

The questions that will be posed and investigated in this research are as follows:

1. What features of mammogram images are considered critical for their pathological assessment and what features have been considered in the literature for characterizing calcification objects in such mammogram images?

2. What feature extraction and similarity measurement techniques perform better in reducing the gap between the human and system semantic interpretation of mammogram image features?

3. To what extent do the existing parallel processing techniques reduce the computational complexity of the implemented feature extraction techniques?

1.5 Contributions of the thesis

This research work makes the following contributions to the retrieval of mammogram images:

1. Identification of the most used and appropriate features used in characterizing microcalcifications

2. Combining existing gradient detection approaches optimally, leading to better detection of microcalcifications in mammogram images

3. It introduces a novel approach for characterizing pathological objects in mammogram images in a manner that improves classification performance and by effect, the retrieval performance of CBMIR systems

4. Development of a parallel model that helps reduce the computational cost of feature extraction as a means of improving the responsiveness of content-based mammogram image retrieval

1.6 Scope of the study

This research work focuses on the retrieval of medical images, and in particular, mammogram images. Full-fledged practical CBMIR systems are composed of several modules that include a user-friendly interface and dedicated storage servers. This study focuses more on the feature extraction process, with light consideration accorded to the other modules. Two parallel computing techniques, multicore and distributed computing, are considered. Besides the Graphical processing

unit (GPU), most parallel computing applications make use of either or both of these two techniques. System evaluation will be done on a standard database to enable comparison with related works. The database used in this study has an accompanying ground truth making the involvement of a medical expert irrelevant.

1.7 Thesis overview

The rest of the thesis is organized as follows: Chapter 2 gives the background on Breast cancer and CBMIR systems, followed by a review of related works. Chapter 3 presents the proposed metholodogy for the detection of microcalcifications. It discusses the integration of various gradient analysis methods for optimizing the detection of microcalcifications. Feature characterization and classification as a means of improving the accuracy and generality of CBMIR systems is covered in Chapter 4, followed by the parallel extraction of features as a means of improving the time efficiency of the process in Chapter 5. A comprehensive discussion of the results, including a comparative analysis of the proposed model in light of related work are presented in Chapter 6. Chapter 7 summarizes the thesis and suggests possible future work.

Chapter 2

Background and Literature Review

The primary focus of this research is the contribution towards the accuracy of feature extraction in Content-based Mammogram Image Retrieval. It also looks at improving the time complexity of feature extraction by leveraging parallel computing techniques. This chapter gives a background of the concepts covered in the research followed by a discussion of related work. A summary of the chapter is given at the end.

2.1 Background

In the following section, we briefly discuss content-based image retrieval and the context of its application in the medical field. This is followed by a discussion on its application to the domain of mammography in Section 2.1.2.

2.1.1 Content-based medical image retrieval

Database Management Systems have to deal with an increasing preference for multimedia information. The current era is experiencing an information explosion, which is facilitated in a major part by advances in technology. CBIR systems are becoming more visible in practical applications as evidenced by systems such as QBIC, Pic Hunter and google image search. This technology is also being incorporated in specialized domains such as diagnostic and decision support systems, face recognition, finger print recognition, etc. CBIR systems in the medical domain are usually referred to as Content Based Medical Image Retrieval (CMBIR) systems [9].

(a) Existing medical infrastructures and organization

Fields such as radiology accumulate image data about patients over a period of time with the aim of building a referential database for future needs. Many modern hospitals have a technological framework for the acquisition, storage, distribution and visualization of information and particularly, multimedia data [2, 9]. The Hospital Information System (HIS) is one such component, that facilitates interchange of information across departments while the Radiology Information System (RIS) provides a computerized system to support operational workflow and business analysis within radiology department through enabling storage, manipulation and distribution of patient radiological images and data. The PACS (See Fig. 2.1) perhaps forms the most significant framework for CBMIR applications due to its core functionality; it is a combination of hardware and software system that provides support for generation, processing, storage, retrieval and presentation of images [2, 9]. It typically is made up of the following components:

- **Imaging modality** this represents the source of image data; it includes the Positron Emission Tomography (PET), Computerized Tomography (CT) and Magnetic Resonance Imaging (Magnetic Resonance Imaging (MRI)) scanners. Can also be a digitizer for converting films
- **Secured Network** these networks employ Virtual Private Networks (VPNs) or Secure Socket Layers (SSLs) technologies to ensure security of the normally sensitive patient data.
- **Workstations** act as interface between the end-users and the system. They are made up of conventional output (e.g. monitor) and input (e.g. keyboard) devices and used for interpretation of images and optional communication of feedback.

- **Archives** provide repository for images and reports

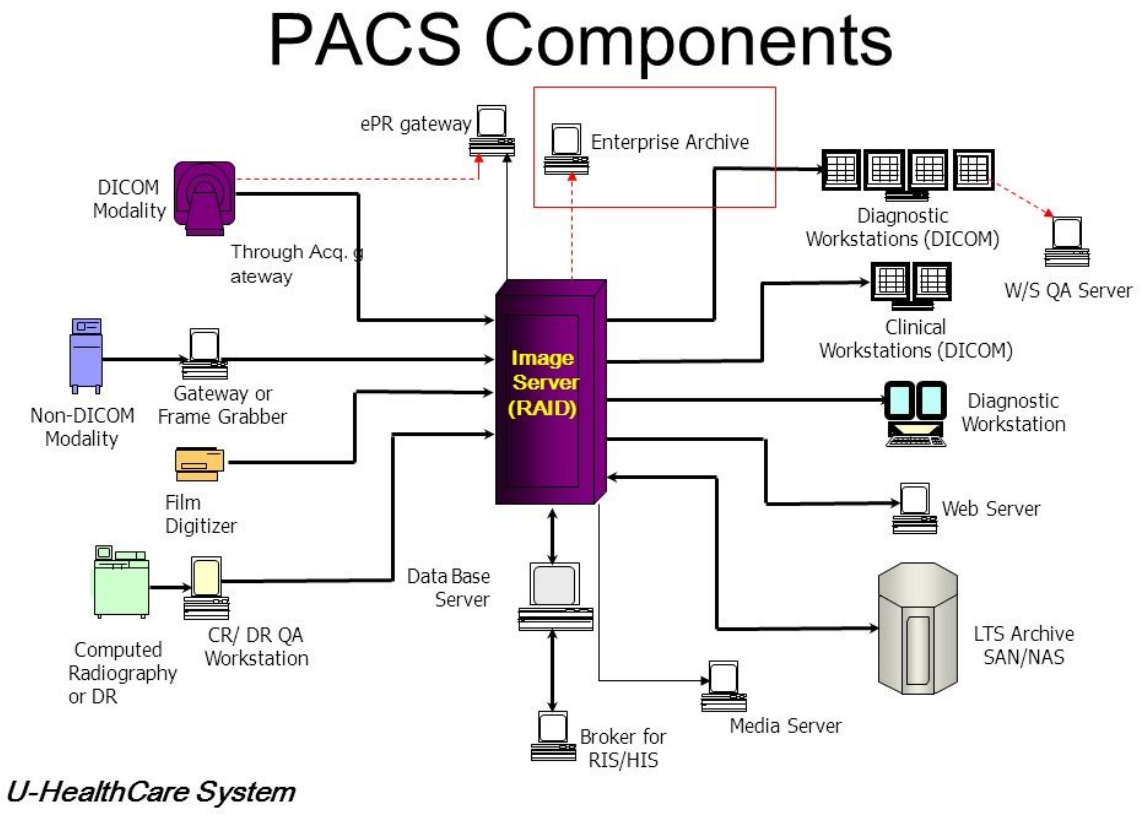


Figure 2.1 PACS showing the sources, sinks and processing flow of medical image data [2]

Industrial components such as the PACS and standards such as Common Object Request Broker Architecture (CORBA) have been developed for simplifying and standardizing information interchange, with the latter being intended for use in distributed object computing [10, 11]. Medical images are commonly based on the Digital Imaging and Communications in Medicine (DICOM) standards for representation and communication protocols. These standards serve to maintain international standards for the communication of biomedical, diagnostic and therapeutic information in medical disciplines that make use of digital images and associated data. Images adhering to such standards are more suitable for automated analysis and open up a wider variety of manipulation and analysis procedures than could be realized under the manual text-based retrieval approach with the traditional image formats.

(b) Traditional text-based queries

The traditional notion of information handling with regards to DBMS commonly made reference to textual data, even when the actual data was multimedia in nature. For instance, while pictorial content has been stored in databases in some applications, access to it has been through textual tags intuitively coined to reference the pictorial data. In radiology departments, screening procedures for patients culminate in X-ray images. These images were annotated using textual tags for future reference. However, manual tagging of pictorial data presents various challenges [12, 13]:

- **Ambiguity of expression among observers** - Inconsistencies are possible at two levels: between the descriptions of the same phenomena by different experts, and between the descriptions of the same phenomena by the same person over a period of time. The latter case occurs where the radiologist acquires more knowledge on the subject making it possible to have a different interpretation of the same phenomena. Even among experts, inter-observer variation rates have been reported up to 80% in some studies [14, 15].
- **Limited dynamism and exhaustiveness of words** - It is almost impossible to capture all the details of an object using words. Furthermore, the saying "an image says more than a thousand words" attests to the fact that more information can be stored and thus conveyed visually than it would textually. This limits the extensibility of the system, since users may want to query certain aspects about the image that may not have been captured by the annotation process [15].
- **Personnel cost** - the annotation option requires comparatively more personnel to implement, given that it is a manual process. This also implies an additional cost to attain the right expertise and experience levels [14].

(c) Search by visual content

The influx of multimedia data poses multi-faceted challenges on the entire information flow cycle of input, output, processing and storage [2, 16–18]. For instance, the system has to cater for input of data using methods other than keying and form entry as is applicable to text-based data. Multimedia data requires more superior processing methods and power to enable real-time response. The storage needs of multimedia data are also greater, their output demands are more graphics-intensive than text. Nonetheless, the benefits of Multimedia data outweigh their challenges since

they allow for more representation, expression and less ambiguity. Given the complex nature of these challenges, a solution design necessitates a paradigm shift with regards to the traditional text-based information handling methods [19].

CBIR refers to the retrieval of a set of images from an image corpus based on visual/pictorial content (otherwise called visual features or simply features) rather than textual annotations/keywords [20–22]. CBMIR is simply the application of CBIR techniques to medical images. Visual attributes have been categorized into three [2, 8]:

- **Primitive features** are features such as color, shape, texture and spatial location. Most algorithms use these features.
- **Derived attributes or logical features** involve some degree of inference about the identity of depicted objects.
- **Abstract attributes** involve complicated reasoning about the meaning of depicted objects.

The CBIR systems were designed with the goal of circumventing the challenges facing text-based queries, as enumerated in the previous section. Other motivating factors for CBMIR systems include: an increasing preference for multimedia information, technological advancement in data generating equipment (such as X-Ray/CT scanners), has led to an abundance of visual data, necessitating a paradigm shift in processing methods to match this new reality [8, 23, 24]. Radiology departments also routinely accumulate image data about patients over a period of time, thus building a referential database for future needs. Most CBMIR systems adopt the model in Fig. 2.2.

The Image Retrieval in Medical Applications (IRMA) project is one of the landmark efforts to improve retrieval tasks over medical image databases, and has motivated many projects in the domain [16]. One of such projects is by Guild et al [25], who presented a general multi-step approach to IRMA. Their study proposes a system based on conceptual and algorithmic separation of the various CBIR steps, which include: Global feature-based categorization, Local feature extraction, feature selection, indexing, identification and retrieval. These steps correspond to semantic layers of knowledge representation, and are sequentially combined. The IRMA concept is related to the blob-world and provides a flexibility that allows for extension by incorporation of new methods such as feature extraction methods. Feature extraction is the capture and representation of information from images; it has a significant bearing on the accuracy CBIR algorithms. CAD systems (including CBMIR systems) can benefit from existing structures and resources in the medical domain [17], such as the HIS, the RIS and the PACS.

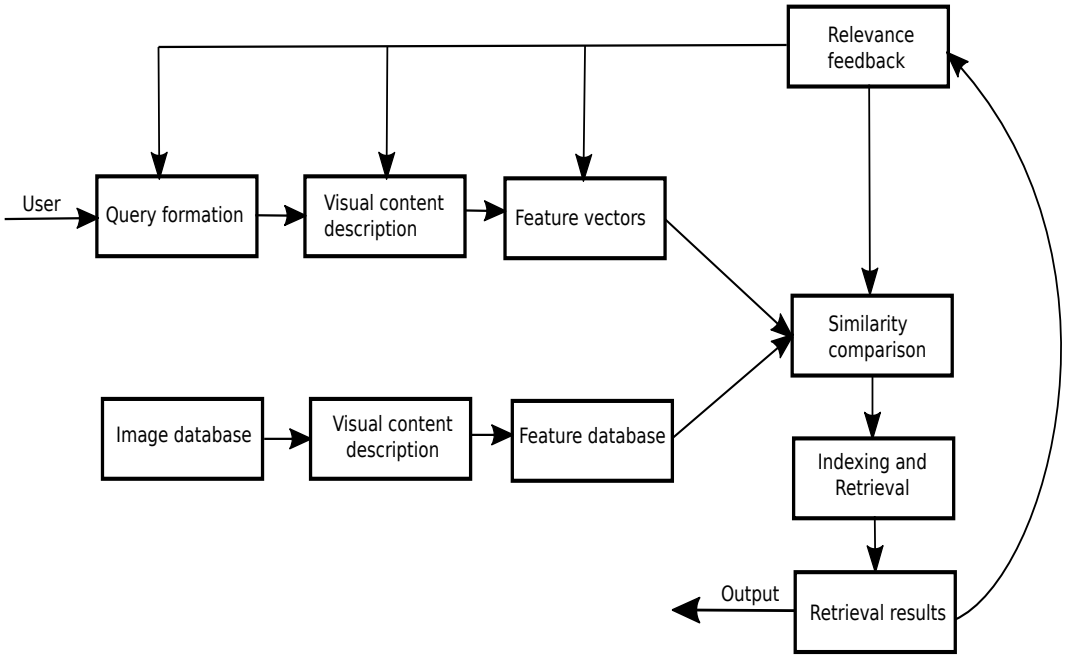


Figure 2.2 General content-based image retrieval scheme. The image database is usually generated in an offline phase

2.1.2 Pathology-based mammogram image retrieval

Pathology-based retrieval is the retrieval of images based on medical properties of objects in those images. Specialization of feature extraction models to certain image subsets enables the inclusion of domain knowledge and assumptions, which significantly improves the performance of such models through customization [26]. Mammography is one domain that has seen targeted research activities, with algorithms tailored to the specific properties of mammogram images. Mammogram tests entail the screening of the breast for cancerous symptoms using low energy X-ray radiation; it offers among the best chances for early detection of breast cancer and has been established to increase survival chances from 20% to 80% [14, 27]. Some alternatives to mammography are Electrical Impedance Spectroscopy (ESI), Infrared imaging, MRI and Ultrasound [17].

Pathologies

Breast cancer is among the leading cancers in women [28, 29]. Calcifications and breast masses are the two most important and prevalent indicators of the disease [14, 30]. Calcifications are calcium deposits in the breast; they form in the Terminal Ductal Lobular Unit (TDLU), which forms the site for invasive cancers. They are classified into 5 levels using the Breast Imaging Reporting and Data System (BI-RADS) scheme, with levels 1-3 being considered benign, and 4-5 being considered malignant [26]. Calcifications are identified based on their site (Fig. 2.3): lobular calcifications are situated in the Acini, while intraductal calcifications are formed in the terminal ducts. Lobular carcinoma has the following additional properties: uniform, homogenous, sharply outlined, mostly punctate/round, occurs in diffuse or scattered distribution; it is mostly considered as benign. Intraductal carcinomas on the other hand vary in size, density and shape (pleomorphic), have fragmented irregular contours with linear/branching distributions; they are considered highly indicative malignancy (classified BI-RADS level 4 or 5) [31].

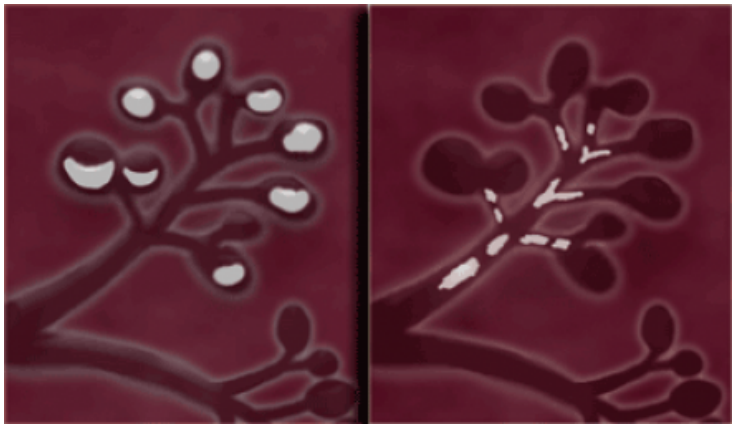


Figure 2.3 Lobular vs. intraductal calcifications. The classification shown here is based no the site of occurence of the calcifications. Lobular calcifications are located in the Acini, while the ductal calcifications are formed in the ducts. [32]

Three factors that influence the diagnosis of calcifications with regards to malignancy are: distribution, morphology and change over time. Changes over time might indicate malignant activity, although such a classification needs to be coupled with the morphological information of the calcifications. Fig. 2.4 shows the various distributions of calcifications, with their classification given in Table 2.1. Basically, the distribution of calcifications is interpreted as follows:

- Scattered/diffused - similar appearing calcifications scattered over breast. Classification favors benign.

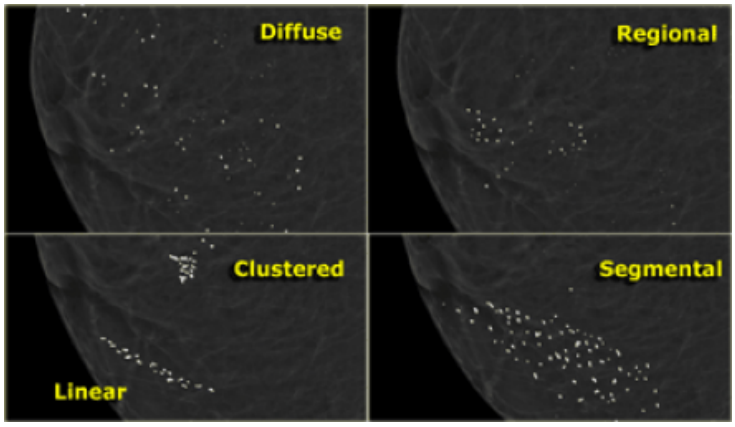


Figure 2.4 Distribution of calcifications. Shape and coverage information can be easily picked from this image and is crucial in their analysis and diagnosis [32]

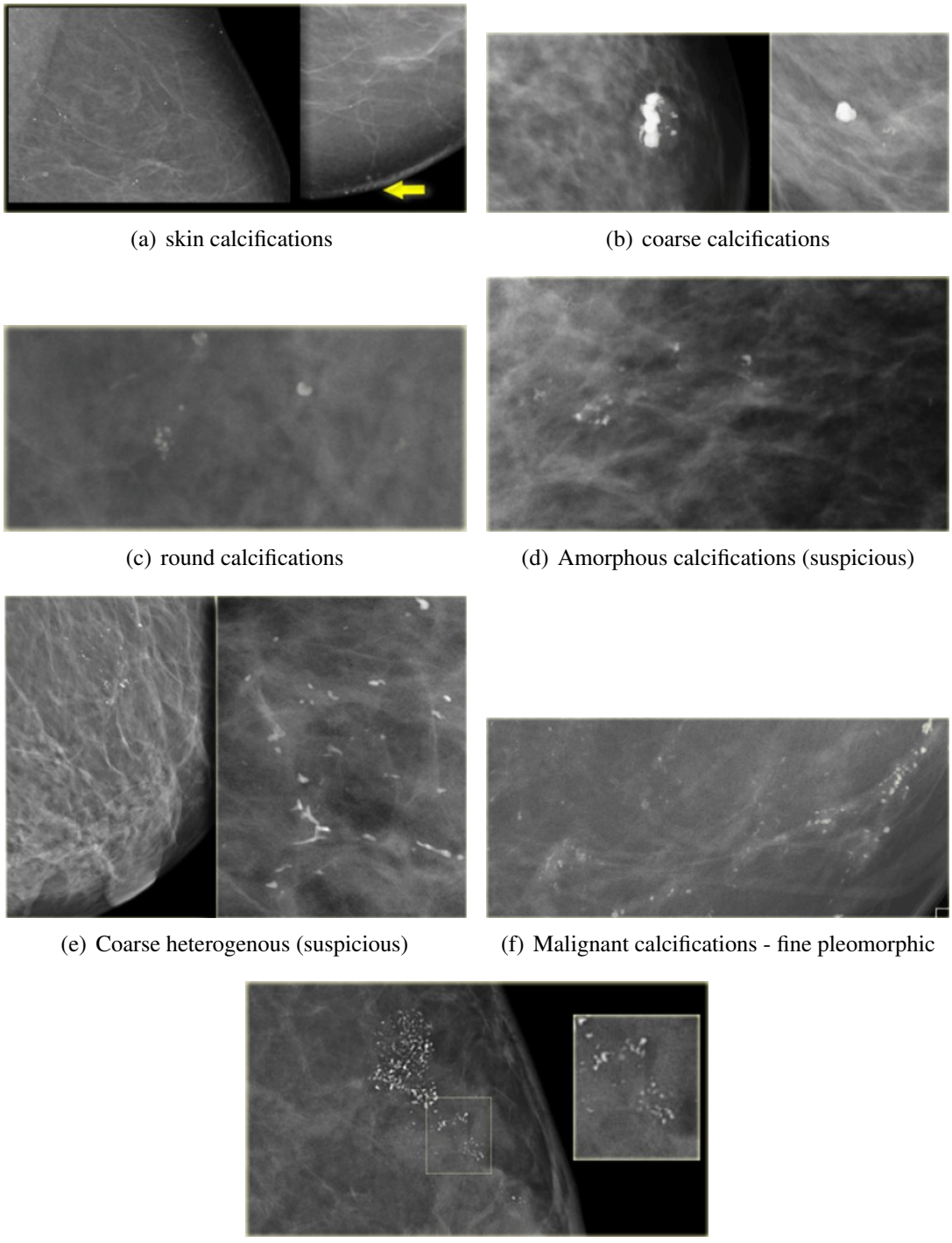
- Regional - scattered in larger volume in breast tissue rather than ductal distribution. Classification favors benign.
- Clustered - at least 5 calcifications in small volume of tissue. Classification could be benign or malignant. Single cluster favors malignant, scattered favors benign.
- Segmental - calcifications situated in ducts or branches of lobes/segment. Implies ductal distribution hence malignancy. Difficult to differentiate with regional.
- Linear - calcifications fill entire duct and its branches. Calcifications with this distribution are suspicious for malignancy.

2.1.3 CBMIR and differential diagnosis

Differential diagnosis is a method of distinguishing diseases by systematic comparison to known, similar pathologies [10, 33, 34]. CBMIR plays an important role of retrieving images of the same form, anatomical region or pathology. In the latter case, it is used as a diagnostic aid tool. This involves the retrieval of a set of similar images to a given query, enabling a radiologist to diagnose the case in hand by referencing the retrieved set of images which, ideally, carry a similar pathology. This is a different approach to Computer Aided Detection tools (CAD) which directly offer a diagnostic suggestion [17]. Survey report findings established that medical practitioners are not receptive to a computerized second opinion that contradicts their own, and are less likely to change their opinion in light of such [35]. It also noted a negative correlation between susceptibility to

Table 2.1 The classification of calcifications by morphology and distribution [32]

Appearance	**Properties**	**Classification**
Coarse	• Large • Popcorn-like	Benign
Round/punctate	$0.5mm - 1mm$	Benign (BI-RADS 2-3)
Amorphous	• Indistinct shape • Small and hazy appearance	Suspicious (BI-RADS 4)
Coarse heterogeneous	• Irregular • $> 0.5mm$ • Small and hazy appearance	Highly malignant
Fine pleomorphic	Variable size and shape	Highly malignant
Fine linear or fine linear branching	• Thin • Linear/curvi-linear irregular shape • Linear/branching morphology	Highly malignant (BI-RADS 5)



(a) skin calcifications

(b) coarse calcifications

(c) round calcifications

(d) Amorphous calcifications (suspicious)

(e) Coarse heterogenous (suspicious)

(f) Malignant calcifications - fine pleomorphic

(g) Fine linear calcifications

Figure 2.5 Mammogram images showing various pathologies related to calcifications. The shape of individual calcifications as well as the clusters is important in differentiating among the types of calcifications. [32]

change and the experience level of the practitioner. The complementary approach of CBMIR-based CAD systems provides a support environment that guides the decision rather than being prescriptive.

The effectiveness of CBMIR in improving the quality of diagnostic decision making has been corroborated in the literature [30, 36, 37]. In one study [37], eight observers were asked to classify twenty unclassified images by their malignancy risk in two phases: first, the observers were asked to label the images without any CBMIR feedback. In the second phase, they were given the same task, but allowed access to eight visually similar images as retrieved by the CBMIR system for every query. The study reported a significant improvement in the classification accuracy of the query image in terms of the A_z value [37] in the second phase, compared to results from the first phase. The authors argue that the better classification performance in the second phase is tied to the perceptually similar cases availed by the CBMIR system. They generalize that differential diagnosis offered by CBMIR systems plays a positively significant role in enhancing the quality of diagnostic decision making.

2.1.4 Challenges affecting CBMIR

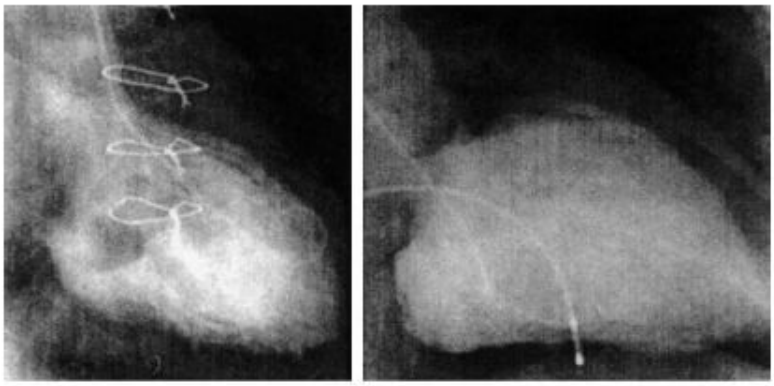
Complete adoptability of CBMIR systems in regular medical practice is hampered by a number of issues [10, 26, 38]. These include inaccuracies of the systems [38] and their limited integrability into existing medical infrastructures such as the PACS [10]. Inaccuracy issues of CBMIR algorithms in certain cases are due to the fact that unlike general images, medical information is imprecise, ill-defined, heterogenous and difficult to obtain automatically from medical images (See Fig. 2.6) [26]. Inconsistencies in interpretation of objects and concepts between algorithms and humans are also another challenge; for instance, since most algorithms work in the feature domain, images have to be transformed to the feature space as an initial task [17, 39]. This transformation potentially results in loss of information, presenting a "gap" between human and machine understanding and representation of the same image. Semantic gap is the most noted gap; it describes the difference between machine and human interpretation of a particular image. Singh et al. [38] discuss up to 12 low-level and high-level gaps that include:

- **Content gap** describes a scenario where algorithms fail to consider the context, and therefore peculiarities, of the image being processed
- **Feature gap** where algorithms fail to capture the significance of objects, which differs ac-

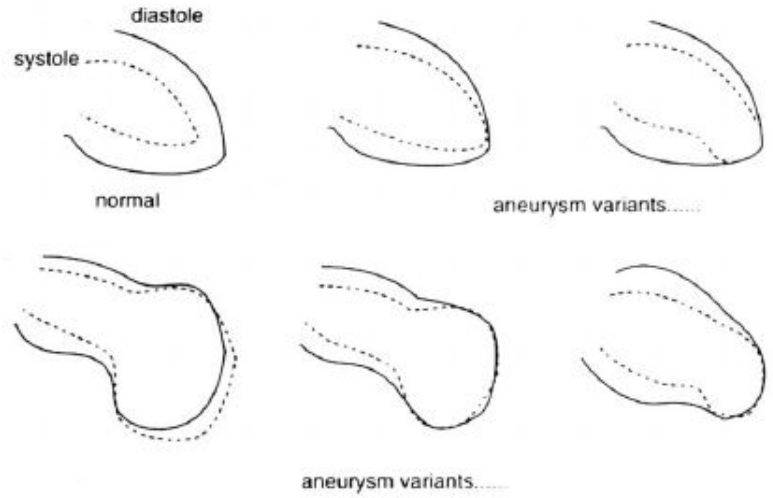
cording to the context. For instance, pathology-bearing regions are more important in medical applications and should be emphasized by algorithms in this domain

- **Performance gap** addresses the lack of algorithms to consider the response time requirements of medical systems
- **Usability gap** refers to failure during the design of the systems to provide clear querying ability, meaningful responses and opportunity for relevance feedback.

Tagare et al [26] list imprecision and heterogeneity of medical images as key issues for medical image retrieval algorithms. Imprecision refers to the inability to precisely articulate concepts in objective and reproducible way (see Fig. 2.6) [30]. This is in contrast to common industrial objects resulting from Computer Aided Design that have established configurational geometry that rarely incurs uncertainty. The heterogeneity aspect of medical images implies that observational findings of certain complex scenarios of medical images can lead to varied inter and intra-observer interpretations. While such varied interpretations might be meaningful in another context, they present an obstacle if a universal formalization of concept is desired.



(a) Cardiac ventriculograms showing aneurysm variants



(b) Illustrations of possible aneurysm variants

Figure 2.6 Candidate Cardiac ventriculograms and illustrations of possible aneurysm variants. This scenario illustrates the difficulty of formalization of the term "ventricular aneurysm"

Medical concepts of diseases usually involve underlying biochemical and biologic processes that have to be factored in resolving heterogeneous situations. Iteration of formalizations is one of the methods proposed as an appropriate solution [26]. This entails formalizing concepts over a smaller database size, increasing the database size and adapting the initial formalization to the new database, repeating the whole process until the whole database is covered. Another proposal involves the comparison of retrieval systems and adoption of best techniques for improving CBMIR systems [26,38]. Good features should also be used to describe images to improve retrieval results.

Multiplicity of features would also bring "new knowledge" that would enhance the distinction of objects. Another factor is the specialization of algorithms to include as much domain knowledge as possible. This allows constraints to be defined thus reducing the complexity space for algorithms.

Research effort is ongoing to improve the accuracy of CBMIR algorithms, to enable them capture the relevant medical characteristics while minimizing irrelevant information. The complexity space is narrowed for CBMIR systems because of domain knowledge which implies targeted application needs [40]. While various levels of satisfaction have been achieved in improving feature extraction, there remains room for improvement. Furthermore, some CBIR algorithms are computationally demanding [27], necessitating the need for design options that will reduce their computational complexity.

2.1.5 Parallel computing

The domain of parallel computing involves leveraging a set of computing resources in a coordinated manner to solve computational problems. This is applicable in situations where the computational load can be divided and solved in parts that are later merged to give the whole solution. Parallel systems can be classed as shown in Fig. 2.7 based on Flynn's taxonomy [41]. Single Instruction Multiple Data computers (Single Instruction Multiple Data (SIMD)) contain an array of processors that execute the same instruction over different data synchronously under a global control unit. The parallel units in a Multiple Instruction Multiple Data (MIMD) are more independent in the sense that they can execute different instructions on different data at a given time. MIMD systems can be further classified by memory organization (Fig. 2.8). Shared memory systems as contrasted to distributed systems infer to the ability of any processor to access any memory location on the system. On a comparative basis, shared memory architectures have the advantage of simpler implementation and design as well as smaller communication cost, but suffer from poor scalability.

Availability of fast networking equipment and powerful but affordable personal computers have led to the pooling together of such resources for coordinated solving of computational problems, in what is called multi-computer configurations. The biggest advantage of such systems is their lower cost as compared to their cummulative processing power. They are also easy to maintain and offer commendable scalability and code-portability. They are however limited by their relatively high latency cost and low bandwidth.

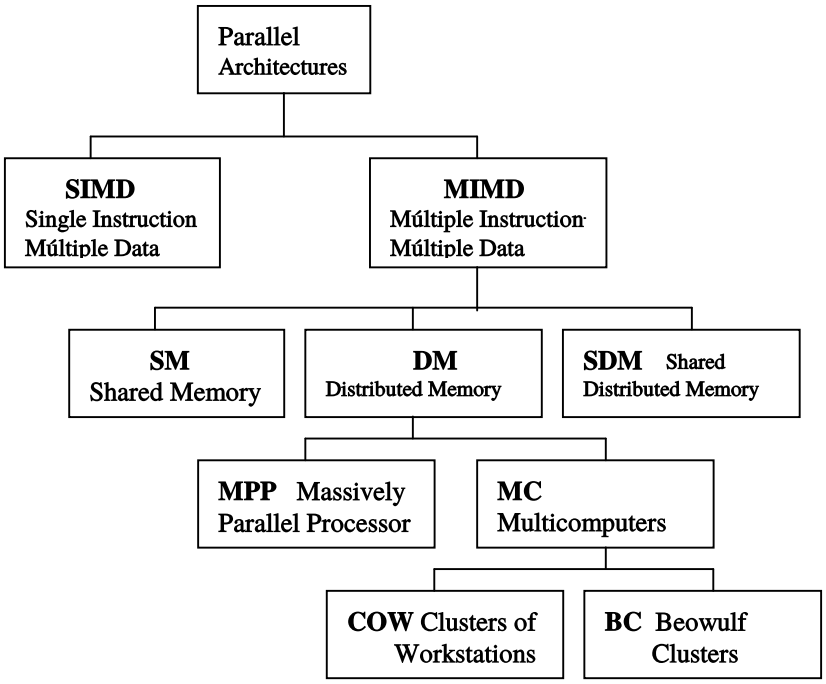


Figure 2.7 Taxonomy of parallel architectures

A major goal in the design of parallel models involves increasing efficiency. Ways of ensuring optimal efficiency are such as drawing a good mapping scheme, minimizing inter-task interactions and ensuring an even load balance across the processors. The following factors affect the design of a mapping scheme,

- **Task generation** - concerns whether tasks are determined/created on the fly (dynamic) or predetermined (static)
- **Task sizes** - involves whether the tasks are of uniform or non-uniform sizes
- **Knowledge of task sizes** - Prior information of task sizes can also influence runtime allocation of tasks.

A chosen mapping technique should aim to keep overheads at a minimum by ensuring minimal inter-process interactions, reduced idleness of processors and avoid excess computations by any given subset of processors.

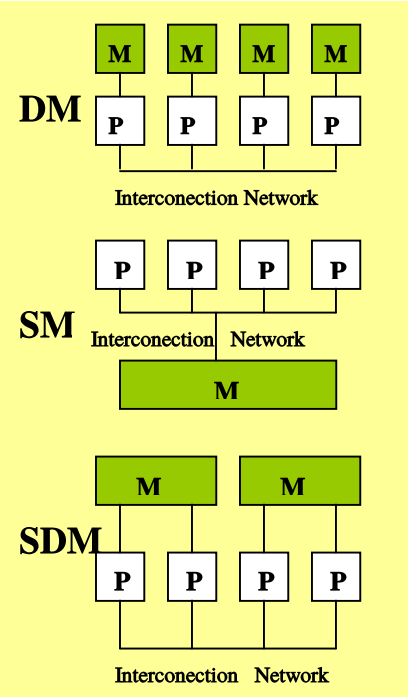


Figure 2.8 Categorization of parallel computers by memory access. This includes the distributed memory (DM), shared memory (SM) and a hybrid approach involving shared and distributed memory (SDM)

2.2 Literature Review

This section looks at research conducted towards improving the accuracy of feature extraction as well as reducing the computational cost of the same. While mammogram image retrieval has different objectives from pattern classification tasks, features that achieve high performance in pattern classification also perform optimally in image indexing [17]. This allows us to consider feature extraction techniques that were used in literature for purposes other than image indexing. Feature selection and classification are also briefly discussed given the relevance and close association to the enhancement of accuracy in CBMIR systems [18].

2.2.1 Feature extraction

Mammogram images generally lack color information, which reduces the scope of applicable features to those that exploit intensity, textural and shape characteristics. Over the last two decades, research has presented various CBIR models for supporting breast cancer diagnosis based on diverse high-level properties such as calcification and mass lesions, the breast parenchyma, asymmetry between breasts, distortion in breast architecture, etc; these are extensively discussed in the survey study by Zheng [42]. For instance, one study suggested age as an effective feature in determination of malignancy on the basis that about 80% of diagnosed breast cancer cases were over 50 years old [27]. In another study, a CBIR model is presented for the retrieval of mammograms based on breast density [39]. It extracts Singular Value Decomposition (Singular Value Decomposition (SVD)) and histogram features, which are used to train a Support Vector Machine (SVM) model. The model is measured on the sole benchmark of average precision, attaining therewith the best score of 82.14% using the polynomial kernel. The authors noted that their model could be improved by considering other crucial information such as features related to lesions as well as appropriate weighting of features. The importance of extracting features directly related to lesions (e.g. masses and calcifications) for CBIR systems is also acknowledged by Kinoshita et al. [43].

The extraction of features related to lesion information for CBIR algorithms has received significant attention in the literature [30, 42]. Salient medical characteristics of these lesions that are crucial for diagnosis are: intensity statistics, shape information and textural information. Generally, features related to microcalcification lesions can be categorized into various categories [27, 30]:

- Individual microcalcification features (perimeter, thickness, area, compactness, elongation, eccentricity, orientation, direction, distance, and contrast)
- Statistical texture features (co-occurrence features, Surround region dependence features (SRDM), Gray level run length (GLRL) matrix features, Gray level difference (GLD) features)
- Multi-scale texture features (Wavelet features-energy, entropy and norm of coefficients, Gabor filter bank features, Laplacian of Gaussian)
- Fractal dimension features
- Cluster features (cluster area, no of microcalcification in area).

(a) Textural features

Texture features are widely used in mammogram image processing to model local spatial variation of the image intensities. They are classified into statistical or structural approaches. Structural approaches consider texture as primitive objects and concern themselves with the arrangement of these primitives. Statistical approaches represent non-deterministic properties that govern distribution and relationship among the intensities. We consider common implementation of texture features in the following sections. Statistical features are extracted in [44] for classifying microcalcifications in mammogram images. Other extracted features include wavelet coefficients, Local binary partition features and median contrast. The features are extracted from a total of 66 ROI's of size 32x32 pixels. Classification is done using the minimum distance and K-Nearest Neighbor classifiers with the accuracy of the classifiers calculated based on their specificity and sensitivity. The work only benchmarked the performance for the classifiers without mentioning the performance of the features themselves. They suggested improvements on the size of the database as well as consideration of more classifiers. Arai et al. [45] also extract statistical features to detect single microcalcifications, as well as microcalcification clusters. They use the Surrounding Region Dependence Matrix and Multi-branches Standard Deviation analysis respectively for the two tasks. The experiment is conducted on 65 ROIs, achieving a classification rate of 70%. Textural analysis is also employed by Tieudeu et.al. [46]. The extracted features are fed to a Neural Network for classification, with the authors reporting good performance of the algorithm according to their results. Further work into texture features can be seen in the work by Wiesmuller and Chandy [47], who use the Normalized Gray Level Aura Matrix (GLAM) to characterize mass and calcification features for retrieval tasks in mammogram databases. Wei et al. [48] extract 132 textural features from 12 GLCMs for retrieval of various breast cancer-related pathologies from the Mammography

Image Analysis Society (Mammographic Image Analysis Society (MIAS)) database. In [49], textural features are used for retrieval of mammograms, by application of a set of Gabor filters. The filtering using Gabor functions is done as a first step to attenuate low-frequencies, the result which is transformed into a probability matrix. This matrix gives the probability of occurrence of the remaining high-frequency components, and forms the basis for computation of six features: Contrast, Angular Second Moment, Inverse Difference Moment, Entropy, Variance and Correlation. A good discussion on the texture features can be found in [50].

Table 2.2 Features related to microcalcification clusters. A cluster is identified where there are at least three microcalcifications within a $1cm^2$ area [51]. Cluster microcalcifications refer to individual microcalcifications forming a particular cluster. Secondary features are those features derived from the primary features, mostly composed of statistical measures. In this table, μ and σ refer to the mean and standard deviation respectively.

Cluster microcalcifications		
Authors	**Primary**	**Secondary**
[27,52,53]	Minimum inter-distance among calcifications	μ, σ
[27,30]	Average intensity	μ, σ
[27,30,51,54]	Area	μ, σ
[27,30,51,54]	Compactness	μ, σ
[27,54]	Fourier descriptors	μ, σ
[27,54]	Moment based measures	μ, σ
[27,54]	Eccentricity	μ, σ
[27,54]	Spread	μ, σ
[27]	Average minimum std. of $r(\theta,l)$	μ, σ
[27]	Average std. of $r(\theta,l)$ at various directions	μ, σ
[27]	Average std of the string at length l, starting from each in calcification object	μ, σ
[52,53]	MCs per unit area	
[51–53]	total number of MCs	
[52,53]	Effective volume (area $\times$ effective thickness)	μ, σ
[51,54]	Density ($\frac{\text{total number of MCs}}{\text{cluster area}}$)	
[51]	Distance from MCs to cluster centroid	μ
[51]	Perimeter (no. of pixels forming MC contour)	σ
[51]	Elongation factor	μ, σ
[51]	Contourlet transform	entropy, correlation, information correlation measures
[53]	Effective thickness	σ
[53]	Second highest microcalcification shape irregularity	

Table 2.3 Listing of extracted features related to the clusters themselves. σ refers to the standard deviation

Cluster features		
Authors	**Feature**	**Description**
[52, 53]	compactness	Roundness of region occupied by cluster
[52, 53]	eccentricity	
[52, 53]	Solidity	Ratio between cross-sectional area and area of MC convex hull
[52, 53]	Moment signature	
[52]	Effective thickness	σ
[52]	Effective volume	σ
[52]	Second highest MC shape irregularity measure	

(b) Shape features

Shape features have also been addressed in literature. For instance, Qi and Snyder [13] exploit shape features for retrieval mammograms with similar lesions. The shape information is captured by eigen values, from which three features are composed: the lengths of the first and second components, and the degree by which their histograms conform to the Gaussian distribution. A similar approach by Felipe et al. [13] is used to capture shape content information of breast lesions using Zernike moments, restricting the retrieval algorithm to those moments that carry the most relevant shape information is found to enhance accuracy of the system.

(c) Wavelet features

Several works have been done in the multi-level representation of mammograms. Rizzi et al. [55] implemented a fully-automated wavelet-based CAD system for the detection of microcalcifications. Two mother wavelets are used in the tasks of microcalcification enhancement and feature extraction, with respect to their individual strengths. The Biorthogonal wavelet (Bior 2.6) was used by the authors for noise removal due to its effectiveness in image reconstruction; it was applied in conjunction with the Donoho thresholding technique. The Haar orthogonal wavelet was used for feature extraction as it does not heavily distort the image. The image was decomposed to a maximum of two levels in both cases. During feature extraction, the decomposition was followed by a binarization process adopting a hard threshold. The wavelet-filtered images were scanned

for microcalcifications by localizing maximum, minimum and crossing-points to form singularity points and passed to a classifier for detecting microcalcifications. In the classification stage, the algorithm checked for clusters by scanning for one or more local maximum/minimum/crossing-point pixels within two windows of dimensions: 6x6, for the decomposition level and 3x3 for the second decomposition level. Experiments reported a sensitivity performance of 98% at a rate of 1 false positives per image (FP/Image), benchmarked on the MIAS database. [56]. The same authors extended their earlier work, adopting the Feed-Forward Artificial Neural Network for classification of microcalcification clusters. The features used as input for the classifier are: minimum diameter, minimum radius, mean radius of clusters and the number of microcalcifications. The Artificial Neural Network (ANN) was trained using 30 images with ten of them containing microcalcification clusters. The maximum iteration for training phase was 1000. Experimental results reported a 98% sensitivity score at 0.65 FP/Image.

Some authors advocate for the Dual-Tree Complex Wavelet Transform over the traditional Discrete Wavelet Transform [57,58]. Preference for this wavelet family is based on its shift invariance, directionality and phase information properties. The property of shift invariance is motivated by the Fourier transform. Structurally, the DT-CWT is a combination of two DWT in parallel forming the upper and lower filter banks, which generate the real and imaginary components respectively. The DT-CWT is constructed as a complex-valued wavelet basis forming a Hilbert pair. This wavelet basis is used to generate 6 wavelets oriented at ± 15, ± 45 and ± 75 degrees. In one study [11], the authors extract fourteen features from the DT-CWT processed image as input to the SVM for classification of microcalcifications. The features include: 3 wavelet coefficients, 9 GLCM-based texture features and two statistical moments features; these features are transformed into a compact set using Principal Component Analysis before being used in the classification by the SVM. The dataset comprises of 50 ROIs of dimension 128x128 from the MIAS database, evenly divided among normal cases and malignant (Microcalcification) cases.

In another study [59], Aquino et al exploit the DT-CWT (Real) for microcalcification detection. This task is divided into four phases: sub-band frequency decomposition, noise reduction, suppression of low frequency bands, dilation of high frequency components and image reconstruction. The denoising technique estimates the noise variance before applying a threshold based on the probability density function of the noise model. The bands containing low frequencies are suppressed to eliminate the mammogram background, and morphological dilation conducted on the wavelet coefficients to enhance any existing microcalcifications. The study was conducted on a set of 15 full images from the MIAS database. The work achieved detection rates of between 20% and 66.6% for all the approaches considered, with the best detection rate achieved with the DT-CWT model. The limitations of the study were reported as the limited accuracy of the model in detecting dense tissues. A potential extension on this work could consider classification of microcalcifications into benign and malignant. The study also did not consider classifier modeling, which might explain the detection performance achieved.

In other works, Balakumaran and Shankar [60] employed a 1-dimensional coiflet transform with multi-scale analysis for the detection of microcalcifications. In this scheme, the transform is applied on the image on a line-by-line basis, both horizontally and vertically up to four levels. The multi-scale analysis involves multiplying two adjacent wavelet detail coefficients to increase singularities in the image while suppressing homogenous regions, effectively amplifying microcalcifications. This is followed by thresholding of the detail coefficients to mark out the microcalcifications. The experimental setup involved ROIs of 100 images from the Digital Database for Screening Mammography (DDSM) database.

In [61], the dyadic wavelet transform is combined with fuzzy shell clustering for microcalcification enhancement and microcalcification cluster detection respectively. To detect the regions containing microcalcifications, the enhanced image is transformed using an undecimated wavelet transform and subdivided into 32x32 blocks, over which skewness and kurtosis are calculated. ROIs are established from those blocks whose kurtosis and skewness values meet a certain threshold. Cluster density and relative shell thickness are subsequently used to determine microcalcification clusters. The algorithm was tested on 112 full images sourced from the DDSM database.

In [62], the Haar wavelet is used for decomposition on 32x32 ROIs that have been manually cropped from MIAS database images. The choice of the ROI dimension is guided by the smallest possible size of microcalcification clusters. The ROIs are decomposed up to four levels, with the input image for each subsequent decomposition being replaced by the approximate component of the previous level. Energy and the infinity norm are extracted as features at each level from all components at each level.

The study in [63] sets out to find the optimal wavelet and decomposition levels for the detection of microcalcifications. The authors carried out a 1D wavelet analysis of two image slice extracts containing microcalcifications using different wavelet families; the decomposition is done up to the sixth level. The data set comprised 40 ROIs that were manually cropped to give the breast region. The experimental results favored the Biorthogonal 2.4 wavelet family at three levels of decomposition. This wavelet basis was therefore used for the second phase, which involved the 2D wavelet decomposition up to four levels of the data set, with the fourth-level zeroing of the approximation coefficients. Local threshold based on statistical moments was applied on the reconstituted images to reduce the number of false positives.

A similar approach is used to find the optimal wavelet [64]. The microcalcification profile is enhanced by setting to zero those detail coefficients with absolute values less than 50% the maximum value, as well as the approximation details of the last level. The wavelet families are compared over 30 ROIs cropped from MC positive images. This study establishes that the performance of

a wavelet family is dependent on the similarity between its function, and the shape of microcalcifications. The Biorthogonal 2.2 at four levels of decomposition gave a better Positive Predictive Value compared to the other families. Classifier algorithms are recommended as further work to reduce the amount of false positives.

In [65], the DWT is used to decompose images up to the third level, followed by extraction of textural features from the HL and LH sub-bands of the third level. Besides being used during feature extraction, wavelet transforms additionally have been proven to be effective in image enhancement. In [58], the DWT was used in combination with morphological filtering to enhance the contrast of mammogram image ROIs, with the authors reporting significant contrast and variance gain in the resultant images.

2.2.2 Feature selection

A significantly high number of features, especially with regard to the number of available image samples, is undesirable in most classification applications; a scenario Bellman [66] called "the curse of dimensionality". The undesirability stems from the fact that many features degrade the performance of a classifier rather than improving it. Not all features have the same discriminatory power, and authors are usually interested in identifying the most effective feature set. Feature selection, also called feature/variable subset selection, is an important step in reducing the semantic gap by removing the influence of irrelevant features. Most selection methods can be classed under variable ranking, subset selection and penalized least squares; they derive from statistics and are based on hypothesis testing frameworks [18]. A smaller feature set also implies reduced computational complexity. Feature extraction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) can also be used as dimension reduction techniques [18]. These techniques however involve re-mapping of features to another space; this leads to loss of information about the original features, the side effect of this is that meaningful rules cannot be extracted from a classifier on the initial features.

Feature selection works in a two-fold manner; selection of a subset of features, followed by evaluation of the selected features by an objective function [66, 67]. The objective function is tied to the feature selection module and can be used to guide further feature subset selection. Feature subset search strategies include: Genetic algorithms, Bi-Directional Search (BDS), Sequential Backward Selection (SBS), Hill climbing, Stochastic search, etc [66]. Objective function techniques can be divided into three; filter methods, wrapper methods and hybrid methods. Hybrid methods are simply a combination of the first two. Filter methods evaluate the effectiveness of

a feature subset by considering individual features based on independent tests such as statistical independence and information theoretic measures. Wrappers use classifiers to test the predictive accuracy of a feature subset through cross-validation and/or statistical re-sampling techniques. A major setback of wrapper approaches is the tendency to overfit to data [67]. They however perform well at establishing dependencies between features.

In other techniques, Felipe et al. [13] used association rule mining to identify the most relevant features for classification between benign and malignant classes. They further incorporate a fractal theory-based technique for dimensionality reduction of the feature vectors. The t-test technique has also been used to rank the relevancy of features [48, 49]. Wei et al. [52] used sequential back selection search procedure to select a 12 dimensional optimal feature subset from an initial 18 dimensional feature set for their mammogram retrieval system. While they acknowledge that it is suboptimal, their choice is based on its ease of implementation. In [48], a statistical t-test is used to enhance the accuracy of process by selecting the most discriminative features from a set of 132 textural features.

2.2.3 Relevance feedback and machine learning

A general trend by authors to bridge the semantic gap is by using machine learning algorithms. For instance, relevance feedback is used to capture and encode information about user preference through feedback rounds, which is used to weight features [68]. The process entails retrieval of initial results, followed by an user-evaluation step, where the user specifies the relevance of the retrieved results. User feedback is usually in form of discrete labels or continuous ratings; it is used to adjust the feature weights to bias subsequent retrieval tasks to the user's preference. The process is usually repeated a certain number of times to further refine the algorithm.

Machine learning has also been used to train classifiers for classification tasks and model parametrization. Instances of classification tasks are such as categorization of images based on malignancy [27]. Some of the classifiers used in this domain include SVM, artificial neural networks (ANN), Genetic Algorithms (GA), Bayes classifier, kernel fisher discriminant (KFD), $k-$Nearest Neighbor (k-NN) clustering, Self-organizing maps (SOMs), Binary decision tree, etc. The SVM and ANN have consistently been ranked among the best classifiers in microcalcification-related applications; a comprehensive discussion and comparison of the two can be found in [27, 52]. Since the SVM classifier is designed to handle a two-class problem, there exist two approaches to deal with classification scenarios involving multi-class problems [39].

- One-against-all – in this approach, every class has an associated SVM. The problem is reduced to classifying a target vector as either belonging or not belonging to this class; this method lumps all other classes into one class, effectively reducing the problem to a binary classification task. Assuming C is the number of classes, the total number of SVMs is $M = C - 1$. All training data is used to train the SVM for one class. A vector is usually assigned to the class whose SVM's discriminant function generated the highest value. This method is used in [39, 51].
- One-against-one – in this approach, an SVM is trained for two classes using only data belonging to the two classes. The number of SVMs required for all the classes can be obtained as $M = (C-1)(C-2)/2$ [39].

Significant factors to consider when using classifiers is to avoid over-fitting and under-fitting to sample data [52]. The number of samples for training and testing should also be sufficient, and effort made to equalize the class distribution as has been discussed in previous sections.

El Naqa et al. [69] implemented a hierarchical learning network for the dual objectives of enhancing accuracy and speeding up retrieval of mammograms. The first stage uses the Fisher discriminant classifier and the Support vector machine (SVM) to coarsely filter through the first batch of related cases. These cases are then fed as training data to the second-stage classifier combination for the modeling of a more refined similarity calculation function. The second stage comprises of the General regression neural network and the SVM classifiers.

Ren [27] investigated the relative performance of the SVM and ANN classifiers in classifying microcalcification cluster (MCC)s. The ANN is configured with 15 hidden layers, training is stopped after about 4000 iterations showing unchanged results. The SVM is based on the RBF kernel with the other parameters established using cross-validation. The study considers two contexts of analysis: balanced learning vs unbalanced learning by using optimized decision making vs not using optimized decision making. In both cases, the database samples for training and testing are divided on a 80:20 ratio respectively. Balanced learning is implemented by over-sampling positive samples. The ANN reportedly outperforms the SVM in both contexts with an A_z score of 0.981 and F_1 score of 0.979 for balanced learning.

An adaptive SVM is used in a cascade topology for classification of a query image based on the results of a pathology-based retrieval framework [52]. Firstly, the standard SVM classifier is used, conditionally followed by the adaptive SVM. The condition is that the first SVM classifier fails to correctly classify samples that are known to be close to the query. The SVM is adapted

by customizing its decision function in light of information from samples close to the query image. The authors conjecture that, if the classifier performs poorly on known samples close to a given query, it has not been well trained on those samples and will not perform well on the query itself; its decision function is thus adjusted using those samples. By imposing a higher penalty on misclassification of samples known to be close to the query, the SVM is modified to be more of a local classifier, akin to the k-NN classifier. Both classifiers have the same parameters, i.e, Gaussian RBF kernel, $\sigma = 2.5$ and $C = 100$.Their method was reported to improve classification results (measured by A_z value) from 0.7752 to 0.8223 with the error rate reduced from 0.31 to 0.26.

Tsochatzidis et al. [51] used an ensemble of SVMs for retrieving mammograms based on pathological similarity along the classes defined in BI-RADS. Each SVM in the ensemble is dedicated to a singular BI-RADS class and tasked to determine membership of a given image to its particular class. The BI-RADS classes considered are: pleomorphic, amorphous, punctate and fine-linear branching. The Gaussian radial basis function is used as the mapping function of the SVM. Instead of considering the sign of the decision function, the distance between the query vector and the nearest support vector are considered instead. In this case $2/3$ of the dataset (containing 87 ROIs) is used for training with the remaining used for testing. The SVM is also used to model similarity in [53]. The parameters used are Gaussian kernel, $\sigma^2 = 1, C = 100$. This retrieval model was found to provide the best results at 72.5%.

2.2.4 Modeling perceptual similarity

A significant number of studies [52,70,71] used the Euclidean distance metric on the primitive feature vectors for measuring similarity. There is a high likelihood that such similarity measurements might not capture the semantics of the image and therefore fail to mirror the user's perspective. One resolution to this challenge has to consider similarity measurements as a classification or clustering problem and apply machine learning techniques [72].

The pivotal concept of user similarity perception modeling with regards to lesions and more specifically, microcalcifications, is presented by El-Naqa et al. [54] as further work on a model presented in their earlier seminal work. In their study, the authors encode perceptual similarity of mammograms by radiologists using the neural network (NN) and SVM classifiers, based on nine microcalcification cluster (MCC) shape features extracted from regions of interest (ROIs). The authors posit that classifiers capture similarity as perceived by human observers more accurately than simple distance metrics. The ROIs forming the image dataset were sourced from a public database and scored by radiological experts specifically for that study. Experimental results

reported a significant improvement in the matching percentage (76.7%) of their learned model against the Euclidean distance metric, even surpassing that of the human observers (66.7%). In a largely similar experimental setup, they expanded on the results of their study by incorporating individual microcalcification features, with the objective of comparing supervised learning (modeled using the SVM classifier with a Gaussian kernel) against unsupervised learning (using Discriminant Adaptive Nearest Neighbor (DANN)) [53]. The results reported a superior matching fraction score for the supervised technique at 72.5% against approximately 64.5% for DANN.

Having demonstrated the viability of classifiers in encoding domain-specific information as briefly discussed in the preceding paragraphs, researchers have also looked at extending/modifying the structure of classifiers in order to customize them to specific problems. This can be seen in the study by Nishikawa et al., [52, 70] where a case-adaptive approach is employed to improve the retrieval performance of their computer-aided diagnosis (CADx) system. Their approach involved retrieving similar mammogram cases for a particular query as a preliminary step using a regular classifier and using the retrieved cases to further modify the decision boundary of classifier. Effectively, the classifier is trained with the new set of retrieved cases in conjunction with the original training set. The computational cost associated with this approach was deemed an issue, which Ho et al. [71] addressed by replacing the decision function of the first classifier (called baseline classifier) with a regularization prior. Apart from achieving a high score according to the Area under the Curve (AUC), the regularized classifier approach [71] resulted in a tenfold reduction in computational complexity.

More recent work on adaptation of the SVM decision function was presented by Tsochatzidis et al., [73], where three SVMs are trained using 90 ROIs from the DDSM database with the task to distinguish breast masses based on three BI-RADS categories. For any given image sample, the authors use the value of the SVMs' decision function rather than its sign as input to a function that calculates what they call the participation value. The three participation values constitute the members of a three-dimensional feature vector that is used for similarity calculations by the Euclidean metric. Their model outscored a state-of-the-art conventional Euclidean-based similarity measurement model by 5.7%. In a subsequent study, their scheme was adapted to microcalcifications covering four BI-RADS categories. Seven shape features and three textural features were extracted to characterize the lesions, with the latter calculated over Contourlet subbands. 87 ROIs extracted from the DDSM database were used for model training and performance benchmarking with the model scoring 60% compared to 52% by the unsupervised CBIR (Euclidean-based model) based on mean average precision.

2.2.5 Algorithm response time efficiency

The reduction of time complexity should be an intrinsic goal of the feature extraction process [3]. Many image processing algorithms involve a lot of computations that might take impractically long periods of time to execute. In a database with many images, feature extraction and similarity comparison are among the most time consuming tasks [74]. Long response times may hinder adoption of these algorithms into practical routines, regardless of whatever other advantages they may offer. For instance, Zheng et al. [75] employed genetic algorithms for selecting the best features and their combinations from a set of 20 features. Computation expense was a major setback with 48 hours taken for an exhaustive permutation search on all feature combinations. It was reported to severely degrade when a feature set with a dimension greater than 11 is chosen. The study estimated an increase to two months for computation if 25 features were to be considered. Ren [27] proposes an improved classification model for MCCs by modifying the classifier's output. This is by determining the classifier's final binary output based on statistical analysis of the classifier's continuous output, a process that is used to set an optimal threshold for binary classification. While the method improves classification results, it increases the computational burden by up to 40%.

Most research aimed at improving CBIR response times is focused on optimizing the efficiency of algorithm design, such as by use of multi-level stages [40,69,76], image downsampling [43] and feature vector dimensionality reduction [13]. For instance, this can be seen in the work by Yang et al. [52], who conditionally use a regular SVM in place of their adaptive SVM to enhance the response time of their algorithm during online query processing. The condition is that the regular SVM correctly classifies similar queries; otherwise the adaptive SVM is used. The advantage in response time is gained by avoiding re-optimization of parameters that would have been necessary if the adaptive SVM were used.

The application of parallel processing techniques in mammogram retrieval algorithms has only started getting attention in recent times, even when such algorithms proved to be computationally expensive [74, 77]. Most applications usually focus on cluster computing architectures, with few targeting multi-core architectures. Parallel processing techniques have been proven to reduce the runtime of tasks, provided that the tasks are coarsely grained. This provides an opportunity for many enhancement and extraction algorithms that have independent sub-operations, which allow for concurrent execution.

Use of grid-based systems has also been found to improve the timeliness of retrieval algorithms in practical applications. Perez et al. [6] deploy on a gLibrary/DRI grid platform their CAD system for diagnosis of six pathological lesions. The system uses four morphological features to train their

Feed-Forward Back-Propagation and Generalized Regression Neural Network (FFBP and GRNN) models on a dataset comprising 100 images taken from the MIAS database. Though they give no actual performance values, the authors acknowledge the contribution of the grid environment in enhancing the reach of their tools. Guild et al. [25] used standard free software and tools to implement a retrieval framework over a database of medical images to satisfactory results. Oliveira et al. [4] also employed the Grid network for retrieval tasks over a medical database comprising of MRI of two anatomical regions: Sagittal knee and axial head.

Wen-hao et al. [74] designed a CBIR system for general images based on the multi-core architecture. The input image is divided into a number of sections based on available processor units before being sent to the cores for feature extraction. The independence of the subimage operations was used to increase concurrency and ensure scalability of the algorithm. Similarity comparison was also done in parallel, with each core carrying out comparisons of the image with a range of database images before sending the results over for aggregation using the parallel merge sort algorithm. The proposed system was established to improve the response time performance in all tasks, with feature extraction being improved by a factor of between 3.10 and 3.28, and similarity comparison being improved by a factor of between 2.82 and 3.75. While the static task partition and assignment scheme in this work might work well in a homogenous hardware environment, there might be challenges in ensuring even load distribution in a non-homogenous environment. The hardware architecture for the experimental runs for this work were not specified by the authors.

Emmanuel et al [78] employed a master-slave model for performing a parallel search and retrieval in an image database comprising of cell, hand and lung images. Wavelet, color and textural features are extracted serially followed by parallel indexing by workers (also referred to as compute nodes) in a cluster containing up to 16 hosts. The slave nodes perform partial indexing and send the intermediate results to the master node which does the final ranking and display of results. Experimental results reported speedup values of between 2 and 10 Notable in that research was that the communication time superceded the total processing time, although the authors did not address it. The partitioning scheme was also static, with the workers allocated equal chunks based on task size as well as the size of the worker pool. Such a non-adaptive task assignment scheme can lead to unequal load distribution in a non-homogenous environment, where the nodes have different processing power.

Researchers have harnessed existing frameworks such the MapReduce framework for speeding up retrieval tasks. One of the popular frameworks for attaining highly parallelized applications is the Hadoop framework. Apache Hadoop is an opensource implementation of the MapReduce paradigm that combines a distributed file system and a programming paradigm called MapReduce for enabling large scale processing of datasets (so-called “Big data”) in a distributed environment, over many inter-connected commodity computers. Its strengths are among others, listed as high

scalability, simplicity and fault-tolerance.

In the literature, Smita et al. [79] present a cluster-based CBIR of mammograms implemented on the Hadoop MapReduce framework. Their model extracts textural, color and shape features in parallel and stores the same in a feature vector on the distributed file system. They however do not present performance data of their model and its implementation details are insufficient. In a similar approach, Jai-Andaloussi et al. [80] applied the MapReduce computing model for extracting features from the DDSM database as part of their CBMIR model. The MapReduce model is used in the offline phase of their project and specifically for the computation of two color-based image signatures using: Bidimensional Empirical Mode Decomposition - Generalized Gaussian Density (BEMD-GCD) function, and Bidimensional Empirical Mode Decomposition - Huang-Hilbert transform (BEMD-HHT). Their model is tested at various scales of the database size. For a smaller datasize ($100 < size < 1000$) their model performs poorly relative to the non-MapReduce model. The performance is equal to the non-MapReduce model at $1000 < size < 3000$ but posts a superior performance at bigger data sizes ($size < 6000$)). Their model thus proved useful only with bigger database sizes.

Graphical Processors Units (Graphical Processing Unit (GPU)) parallelization is also increasingly being used in embarrassingly parallel problems, usually employing the data parallel model. For instance, Kuldeep et al [81] used GPU parallelization for the extraction of features in a CBIR system on a dataset containing MRI, CT-scan and X-ray images, to a reported average speedup of $30x$. Similarly, Heidari et al. [82] achieve a $6.305x$ speedup using their GPU-based model to extract color features for a CBIR system. While the speedups achieved in both works are impressive, such systems would possibly benefit from combining of the other parallel technologies [74] (multiple nodes in a distributed architecture as well as multithreaded computing).

Efficient feature characterization is critical to CBIR CAD-based systems [83, 84] and is still an active research area. Much work is still needed on the characterization of features in order to improve the accuracy of CBIR systems [83]. Similarity modeling using classifiers has demonstrated its viability over simple distance measures as has been discussed in the preceding paragraphs. However, to the best of our knowledge, none of the previous work has considered using statistical descriptors based on the classifiers' decision functions. This work aims to further explore this idea by deriving statistical features from classifier scores as a means of improving the accuracy of CBIR-based CAD systems in the domain of breast cancer diagnosis.

Parallel processing offers a significant chance at improving the response time of CBIR systems, without necessitating drastic redesigning of algorithms to make them more time efficient as is inevitable in techniques such feature dimension reduction or feature space transformation.

While some research effort has been done to significant progress, it cannot be deemed conclusive and it remains to be seen what other approaches and models can offer at improving the efficiency of parallel systems [78]. For instance, the research discussed previously relies on existing frameworks such as the Hadoop framework (Smita et al. [79] and Jai-Andaloussi et al. [80]) for parallel computation of tasks. While such frameworks accord an extensively tested environment and set of techniques, which standardizes and simplifies the development and implementation of parallel models, the same advantages can be a limitation, as it makes them prescriptive by confining one to the availed functionality as is characteristic of frameworks. In a comprehensive analysis of the challenges of MapReduce frameworks, Grolinger et al. [85] point out the "limited optimization of MapReduce jobs". The authors state that the MapReduce paradigm does not always sufficiently provide a means of describing every computation problem and that the model does not natively support the composition of jobs.

The relatively limited research focus - especially when the domain is narrowed to retrieval tasks on mammogram image databases - as well as potential for improvement by considering more parallel models in the parallel extraction of mammogram pathological features forms the motivation for this work. Based on promising results by recent research incorporating parallel computing into CBMIR systems and other related areas as discussed in the preceding paragraphs, this work aims to leverage the cluster and multi-core architectures in a homogenous distributed computing environment as a means of speeding up the extraction of geometric and textural features. The focus of the work is on reducing the communication overhead by ensuring an optimal task assignment model. Parallelization is targeted only at the extraction of features, which is among the expensive tasks in the CBIR system. To the best of the author's knowledge, the parallel model presented in this thesis is novel.

2.3 Conclusion

Towards arriving at diagnostic decisions, radiologists analyze mammogram images for presence of pathological objects, which may be indicative of certain diseases such as cancer. The nature of the pathological objects, if they exist, might give out additional information such as severity and extent of the disease. In difficult-to-diagnose cases, radiologists might make reference to historical cases when processing a given case. Furthermore, radiologists usually consider previous cases when monitoring the trend of the disease for a particular patient. CBIR-based Computer Aided Detection/Diagnosis (CADe/x) systems can aid radiologists by providing diagnostic decision support for medical image interpretation in this scenario, by retrieving cases that are "pathologically similar" to the case in hand.

The adoption of CADe/x systems in widespread practical routines is hampered by their low accuracy rates, relegating them to experimental research domain. The concerns about these algorithms include their failure to bridge the semantic gap. This means that the features used for characterization fail to accurately capture the pathological objects as desired by radiologists. Studies have highlighted the need for the system response to be "meaningful, timely and sensitive to the image acquisition process". Researchers have attempted to explore various feature sets to improve the accuracy of characterization to various levels of success. Nonetheless, the need still remains for effective combination of features to give feature extraction models a multi-lateral perspective and understanding of the image effectively increasing their accuracy of representation.

The response time of CBIR systems is an important factor that had not been commensurately addressed. The retrieval problem is also computationally expensive, especially where a large database is involved. Additionally, some algorithms are intrinsically computationally demanding; this can be compounded with further scaling of the image database, which is an inevitable trend. This poses a challenge since diagnostic decisions need to be made fast in order to allow early commencement of treatment if need be. Algorithmic workarounds such use of dimension reduction techniques have been used to reduce the computational burden. Such techniques however can compromise the quality of the feature vector if used in a trade-off manner, solely for the purpose of reducing the cost of the task. Advances made both in terms of computing power and cost in the domain of parallel processing offer a better opportunity at reducing the computational cost through concurrent processing. The application of parallel is only picking up, and there exist opportunities for improvement. This thesis presents a model that contributes to the improvement of the retrieval accuracy through an efficacious feature combination, as well as the reduction of computational cost of feature extraction using parallel processing. The next chapter presents the proposed methodology with regards to maximizing the detection of microcalcifications, which is an important step in the overall performance of a CBIR system.

Chapter 3

Proposed methodology for microcalcification detection

3.1 Introduction

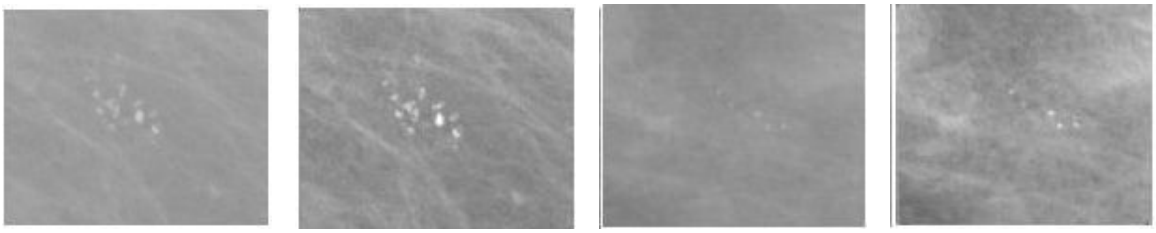
Mammography is a popular approach in radiology that employs safe levels of X-ray radiation to highlight suspicious regions in the breast [86]. One of the common breast cancer indicators visible in a mammogram are calcifications; these are traces of calcium deposits in the breast ducts and lobules, which are visible in a mammogram as high intensity spots with spatial dimensions of between $0.05mm$ and $1mm$. They generally fall into two categories: macrocalcifications, and microcalcifications. Macrocalcifications are relatively larger and rounder than microcalcifications; they are usually considered benign. Microcalcifications are of greater interest to researchers since they are highly indicative of breast cancer, especially if they occur in clusters [87].

Common challenges in microcalcification detection in the literature presented above include undesirable properties of mammogram images such as, poor contrast, noise and aritificial objects. The density of the breast contributes in difficulty of microcalcification detection, with mammograms containing glandular and dense-glandular tissues proving relatively difficult to check for microcalcifications than those with Fatty tissues. This chapter discusses the combination of different feature maps as a means of increasing the detection sensitivity of microcalcifications in mam-

mogram images. Fig. 3.1 presents the block diagram of the proposed methodology for detecting microcalcifications, with a discussion of the phases/stages given in the subsequent sections.

3.2 Preprocessing

While their profile is clearly defined in terms of intensity, size and distribution, microcalcifications commonly are indistinguishable from their backgrounds in many mammogram images (Fig. 3.2), a factor that highly contributes to missed detections [88]. They have various sizes, shapes and distributions, low contrast and are closely connected to surrounding tissue, which makes it difficult to detect them, especially in high density tissues. Other obstacles are artificial artifacts (such as labels and markings) and noise introduced by instruments during image acquisition [16, 27, 51]. Inaccuracies by detection algorithms are also caused by film emulsion errors, digitization artifacts, anatomical structures (fibrous strands, breast borders, hypertrophied lobules). The acquisition environment can also vary, implying differences in illumination that may affect contrast-based algorithms.



(a) Original image sample 1 (b) Enhanced version of original image sample 1 (c) Original image sample 2 (d) Enhanced version of original image sample 2

Figure 3.2 Enhancement applied to two sample images. It is visually evident that the calcifications are easily noticeable in the enhanced versions of the images.

Preprocessing aims to improve the distinction between salient objects and their background [14]. Common preprocessing techniques in this domain are such as [89]: noise suppression, contrast enhancement [27], gray level manipulation, background removal, interpolation and magnification, edge crisping and sharpening, etc. The preprocessing techniques could be manual [39] or automated. While the aforementioned are common techniques used in preprocessing, there is no rigid class of preprocessing techniques, as it sometimes depends on the context of application for

any particular technique. For instance, features could be extracted for the purpose of preprocessing an image, in which case the feature extraction technique will be classed as a preprocessing step. This is usually the case with multi-scale analysis techniques such as wavelet transforms. The performance of preprocesing techniques is usually measured based on the signal to noise ratio (SNR). In this work, the input image is preprocessed by subtracting the mean intensity from all pixel values as

$$I_{out} = I_{in} - mean(I_{in}) \tag{3.1}$$

where I_{out} is the output after the operation and I_{in} is the input (original) image.

3.3 Wavelet analysis

Wavelet transforms have been used for multi-resolution analysis of microcalcifications in mammogram images. The transformations involve modelling the input signal as a superposition of wavelet basis functions, allowing the analysis of singularities and discontinuities in the signal. Unlike the Fourier transform which only allows the modelling of frequency information, wavelet transforms allow simultaneous signal space and frequency localization. This localization property allows to isolate noise, edges, or other discrete objects by filtering out the corresponding frequency. These properties make wavelets suitable for applications targeting localized high-frequency events or scale-variable processes.

The multi-resolution analysis property of wavelet functions is achieved by translating and dilating the wavelet mother function; as illustration, the wavelet family of a Discrete Wavelet Transform is generated as

$$\psi_{j,k}(x) = \frac{1}{\sqrt{a_j}}\psi(\frac{x-b_k}{a_j}) \tag{3.2}$$

Where a_j and b_k are the translation and scaling parameters respectively.

The orthogonal wavelet comprises of two continuous-valued functions: the scaling function $\phi(x)$ and its corresponding wavelet function $\psi(x)$, which constitute an orthonormal basis of $L^2(R)$. By the orthogonality property of these functions, the wavelet $d(j,k)$ coefficients, and approximation coefficients $c(j,k)$, can be obtained as an inner product of the input function with the wavelet and scaling functions respectively as

$$d(j,k) = \langle f, \psi_{j,k} \rangle \tag{3.3}$$

$$c(j,k) = \langle f, \phi_{j,k} \rangle \tag{3.4}$$

The one-dimensional wavelet transform f of an input function is therefore obtained as

$$f(x) = \sum_k d(j,k)\psi_{j,k} + \sum_k c(j,k)\phi_{j,k} \tag{3.5}$$

The above transform can easily be extended to the two-dimensional space. At implementation level, two-dimensional wavelet transforms analyse images by performing a one-dimensional analysis on rows first, followed by columns to generate three detailed sub-images HH, HL, LH, and one approximation sub-matrix LL. The synthesis operation simply reverses the process, integrating up-sampling if down-sampling was done during the analysis stages.

Since microcalcification objects are described as localized intensity spikes with spatial dimensions of between 0.05mm and 1.0mm, they can be analysed in the wavelet domain. As a matter of fact, the Wavelet transform allows for denoising of the image in addition feature extraction. The common approach is to transform the image using wavelet filters, process the resultant wavelet coefficients and synthesize the wavelet representation to give back the processed signal. The choice of the particular wavelet family, depth of decomposition [23], wavelet domain processing techniques, resampling and other related processes is the subject of research activity [23]. This work considers the Daubechies 1 (Db1) family for detection of microcalcifications; this is based on its relatively good performance in related works [62–64]. The following steps are taken during wavelet analysis,

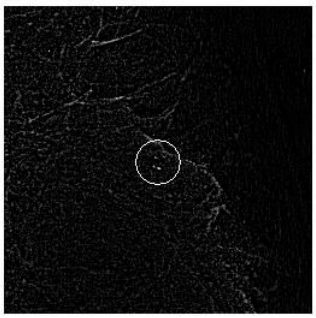
1. Perform a one-level decomposition of the Input Image I
2. Nullify the Approximation coefficients
3. Reconstruct the image using all first level Coefficients to give the detail-enhanced image I_e
4. Rescale the intensity range of the enhanced image to that of the original image to give the final image I_{wv}

3.4 Gaussian/Median filtering

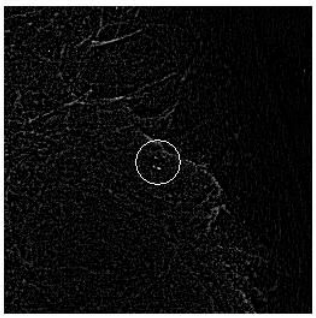
The Median filter falls under order-statistics, non-linear filtering approaches, and is preferred for its preservation of edge information, which is lost in applications involving linear smoothing filters [90]. This makes it popular in applications that target denoising of images which contain salt and pepper noise [91]. Three square kernels are chosen to approximate the varying spatial extent of calcifications. The median filter response map I_{Med} is obtained using the following linear combination:

$$I_{Med} = \frac{2}{3}M_5(i,j) + \frac{1}{6}M_7(i,j) + \frac{1}{6}M_9(i,j) \tag{3.6}$$

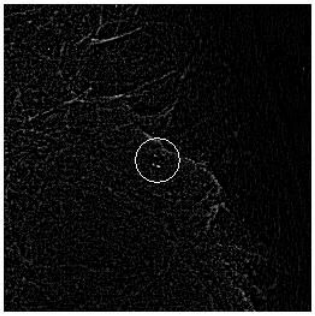
$M_d(i,j) = Med_{d,d}i,j$ denotes the square median filter with a spatial extent of $d \times d$. Calcification objects in the context of the images in this study are approximated at between 3×3 and 9×9 pixels, considering the digitization parameters of the MIAS database [92], and that their spatial dimension is reported to be between $0.05mm$ and $1mm$ [87]. This basis also informs the range of spatial dimensions used in the spatial-domain filters described in the following sections. From experimental runs, the larger spatial filters proved to amplify curvilinear structures at the expense of calcification objects, which led to segmentation challenges. Fig. 3.3 exemplifies the Gaussian response maps for the smallest and largest kernels (the pattern is similar with the Median and Finite Impulse Response (FIR) filters for corresponding kernel sizes). For this reason, the output of the larger dimension filters is weighted lesser than that of the smaller ones. The resultant image is subtracted from the original unfiltered image for isolation of Calcification-like objects.



(a) Response map for the 3×3 Gaussian filter



(b) Response map for the 5×5 Gaussian filter



(c) Combination of the 3×3 and 5×5 Gaussian filters

Figure 3.3 Response map for the Gaussian filters, followed by their combination

The Gaussian filter was chosen for its similarity to the profile of Calcifications, which implies that it generates a strong response in the presence of calcification-like objects. The Gaussian function used to generate the kernel values Eq. 3.7 is defined as

$$G_s(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.7}$$

A higher σ value increases the degree of attenuation. For the σ value, a set of 30 values

linearly spaced between 2 and 5 were evaluated, with the best performance achieved at $\sigma = 2.9$. The filtering results are combined with the following weighting (using scalar multiplication) to define the Gaussian-enhanced image I_{Gaus} as

$$I_{Gaus} = \frac{2}{3}G_5(i,j) + \frac{1}{6}G_7(i,j) + \frac{1}{6}G_9(i,j) \tag{3.8}$$

where,

$$G_d = I * Gaus_d \tag{3.9}$$
$$(i,j) \in S \text{ and } S \subseteq N^2 \tag{3.10}$$

$Gaus_d$ denotes the Gaussian kernel with a kernel size of $d \times d$.

3.5 Finite Impulse Response (FIR) filter

This step uses three Laplace operators described in [93] to enhance microcalcifications (See Fig. 3.4).

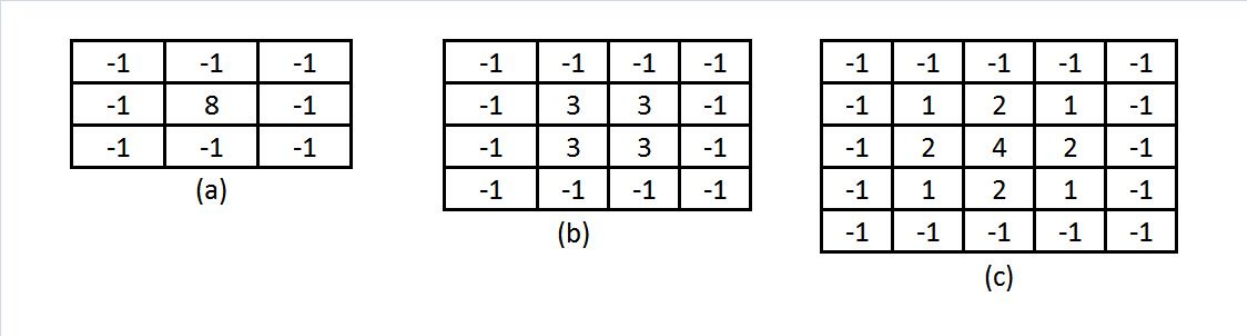


Figure 3.4 Finite Impulse Response filter kernels, showing the 3×3, 4×4 and 5×5

The operators have dimensions 3×3, 4×4 and 5×5. The response map is obtained by filtering the input image with the FIR filters. The three resolutions were empirically chosen to cover the

various ranges in size of calcification objects. The final filter response I_{FIR} is obtained using the hadamard product of the individual FIR filters as

$$I_{FIR} = F_3(i,j) \circ F_4(i,j) \circ H_{5,3}(i,j) \tag{3.11}$$

where,

$$F_d = I * FIR_{d,d} \tag{3.12}$$

$$H_{5,3} = F_5 - F_3 \tag{3.13}$$

$$(i,j) \in S \text{ and } S \subseteq N^2 \tag{3.14}$$

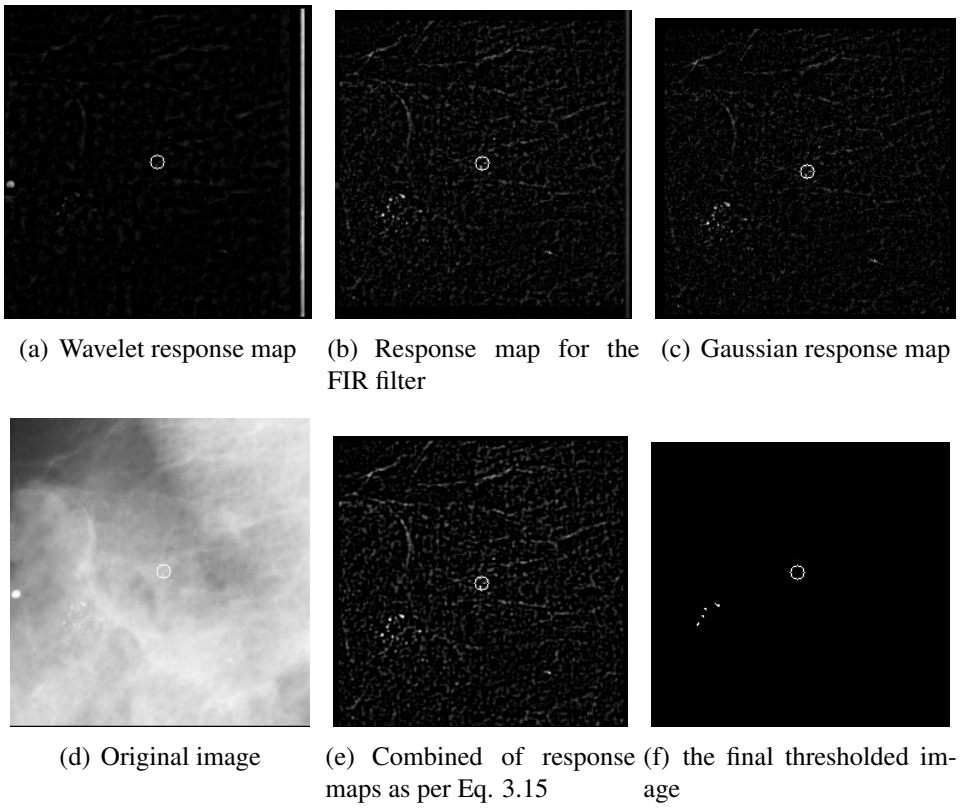
$FIR_{d,d}$ is the FIR filter with the kernel dimensions $d \times d$. The filter kernel F_5 amplifies curvilinear structures, which makes it difficult to segment calcifications. Eq. 3.13 is thus used to diminish the strong curvilinear response.

3.6 Combination of filter responses

Each of the above filters provides a likelihood map for every pixel. Those pixels with high intensity values imply a high probability for presence of a microcalcification object. The image results from the convolution operations with the discussed filters are combined using the hadamard product to compute the final calcification-enhanced image I_e as

$$I_e = I_{Med} \circ I_{Gaus} \circ I_{wv} \circ I_{FIR} \tag{3.15}$$

Where,
I_{Med} - is the median-filtered image,
I_{Gaus} - is the Gaussian-filtered image,
I_{wv} - is the wavelet-filtered image,
I_{FIR} - is the resultant image after applying the finite impulse response filter,
$\circ$ - represents the hadamard product



(a) Wavelet response map (b) Response map for the FIR filter (c) Gaussian response map

(d) Original image (e) Combined of response maps as per Eq. 3.15 (f) the final thresholded image

Figure 3.5 Image results for individual filtering processes as well as their combination and subsequent thresholding

The above operation significantly reduces smaller pixels values with an inverse effect on pixels having bigger intensity values. Furthermore, pixels flagged as suspicious by all the filters are significantly boosted while those that are not flagged by either of the operators are almost nullified. The simple scalar multiplication in Eq. 3.15 was found to significantly reduce the effect of pixels that did not elicit a strong response from all the filters. This had the intended effect of nullifying artifacts that were unique only to a smaller subset of the filters. An illustration of this process can be seen in Fig. 3.5.

3.7 Breast background artifact removal

All the filters used in this work provide a strong response along the boundary between the breast area and the image background in some of the images, which is an undesired side effect (An example can be seen in Fig. 3.5 (a) on the right side). This boundary artifact is eliminated by thresholding the filtered image using the mean value, followed by binary erosion (using a circular structuring element) of the image foreground as shown in Algorithm 1,

Algorithm 1 Remove Breast Area/Image Background Boundary Artifacts

1: **function** REMOVEBBARTIFACT(I_{orig}, I_{proc})
2: $t = mean(I_{orig})$ ▷ The threshold is the global mean
3: $s = GetDiskStructingElement(5)$ ▷ Disk structuring element of dimension 5×5 for erosion
4: $mask = Thresh(I_{orig}, t)$ ▷ Threshold I_{orig} using t
5: $mask = Erode(mask, s)$ ▷ Erode Binary image using structuring element
6: $I_e = mask * I_{proc}$ ▷ Enhanced image is a product of the mask and the processed image
7: **return** I_e
8: **end function**

3.8 Removal of small objects and linear structures

Region size and eccentricity are used as criteria for the removal of small objects as well as linear structures. Objects having an area $A < 3$, as well as those having an eccentricity value $E > 0.98$ are removed as shown in Algorithm 2,

An eccentricity value of $E = 0$ represents a spherical shape with $E = 1$ being a linear object. Some noise objects were found to be near spherical, with high eccentricity values. Others were noted to have small areas of between one and two pixels. These two factors guided the choice of the eccentricity and minimum area thresholds.

Algorithm 2 Removal of Small objects and Linear Structures

Require: $I(x,y) \in [0,1]$ ▷ I is assumed to be a binary image

```
function REMOVEOBJECTS(I)
    A_thresh = 3
    E_thresh = 0.98
    Regions = GetRegions(I)          ▷ Regions represents all image objects
    for all r ∈ Regions do
        A_r = GetArea(r)             ▷ Retrieve the object's area value
        E_r = GetEccentricity(r)     ▷ Retrieve the object's eccentricity value
        if A_r < A_thresh or E_r < E_thresh then
            I(r(:)) ← 0              ▷ Nullify pixels of region r
        end if
    end for
    return I
end function
```

3.9 Thresholding

Entropy information is used to determine the optimal threshold value for segmenting microcalcification objects from the background. This work customizes the Tsallis entropy thresholding technique discussed in [94] over a two-class problem, which is the solution to the following function:

$$f(t) = Argmax\left[S_q^A(t) + S_q^B(t) + (1-q) * S_q^A(t) S_q^B(t)\right] \tag{3.16}$$

$$S_q^A(t) = \frac{1 - \sum_{i=0}^{t-1}\left(\frac{P_i}{P^A}\right)^q}{q-1}, P^A = \sum_{i=0}^{t-1} P_i$$

$$S_q^B(t) = \frac{1 - \sum_{i=t}^{L-1}\left(\frac{P_i}{P^B}\right)^q}{q-1}, P^B = \sum_{i=t}^{L-1} P_i$$

where, q is the entropy index, S the measure of entropy, $L-1$ the maximum gray level in the image region and P_i the probability of gray level i. This work considers segmentation of two classes: The breast background (A) and microcalcifications (B). The image background is ignored, which means that the search begins with $t > 0$, with t taking on the value of the minimum non-background pixel value. Before threshold calculation, the image is transformed to have gray levels in the range $I(x,y) \in [0,255]$.

3.10 Conclusion

This chapter discussed the automated detection of microcalcifications in digital mammogram images. It also investigated the discrimination between Malignant and benign subclasses of microcalcifications. To this end, a set of gradient-based filters were optimally integrated to amplify microcalcifications followed by postprocessing to reduce the number of false positives. Each of the filters have their unique strengths and side-effects in detection of calcification-like objects; the essence was to combine them optimally to highlight their strengths while canceling their side-effects. An entropy-based threholding technique was finally used to determine an adaptive threshold value for isolating calcification objects from their background. The sole use of gradient-based filters is prone to highlighting noise objects, which leads to a high number of false positives. Classification is commonly used to train algorithms to differentiate between noise objects and genuine calcification objects. The next chapter presents the proposed methodology that incorporates feature extraction and classification for alleviating the negative effect of noise objects that have a similar profile to calcification objects.

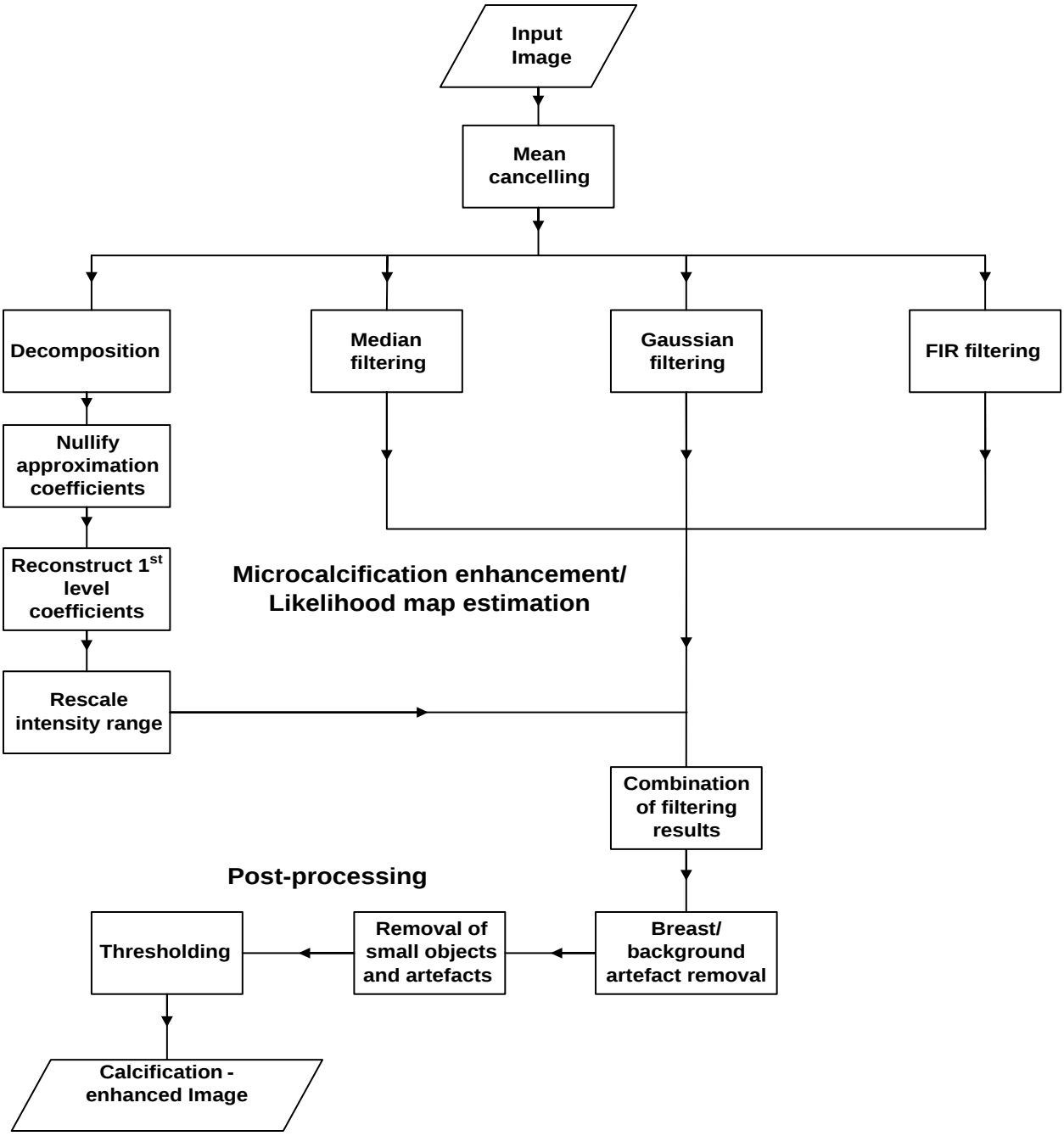


Figure 3.1 Block diagram of the proposed microcalcification detection method. Mean canceling involves subtracting of the mean from the image and constitutes the sole pre-processing activity in this phase. Since detection is the major objective of this chapter, pre- and post-processing in this context are context-specific, and are described as those activities that prepare the image for the filtering stage (hereby labeled microcalcification enhancement/likelihood map estimation) which constitutes the detection stage of the process.

Chapter 4

Mammogram Image Content-based Retrieval

4.1 Introduction

CBIR systems can be used as a second opinion for radiologists by availing pathologically similar past mammogram cases , thereby improving the confidence of the diagnosis [36,39]. An important distinction between CBIR systems and traditional text-based query systems is their use of visual features for queries rather than textual annotations. The efficiency and efficacy of medical CBIR systems - also called Content-based Medical Image Retrieval (CBMIR) systems - strongly relies in part on the features selected for representing the salient high level medical characteristics of the image [95]. Furthermore, the choice of a corresponding method for measuring similarity is crucial in reducing the semantic gap, defined as the difference in high-level interpretation of images by humans as compared to the low-level understanding of the same by algorithmic models. The efficiency and efficacy of pathology-based CBIR systems in providing accurate results to radiologists is a critical factor in their acceptance in regular medical routines [83]. In this chapter, a CBIR model is presented, that demonstrates improved retrieval performance of mammograms based on microcalcifications as the pathology. Microcalcifications, besides breast masses, are the most important lesions in the diagnosis of breast cancer.

4.2 Image retrieval schematic

The proposed CBMIR schematic is presented in Fig. 4.1. The model takes as input a binary image containing probable calcification objects as the foreground and its original gray level version. The term “calcification” or “calcification object”, especially in the early sections of this chapter, is loosely used to refer to the foreground pixel regions in the binary region of interest and does not necessarily mean that the object has been established as a true calcification. The subsequent processes are discussed in the following sections.

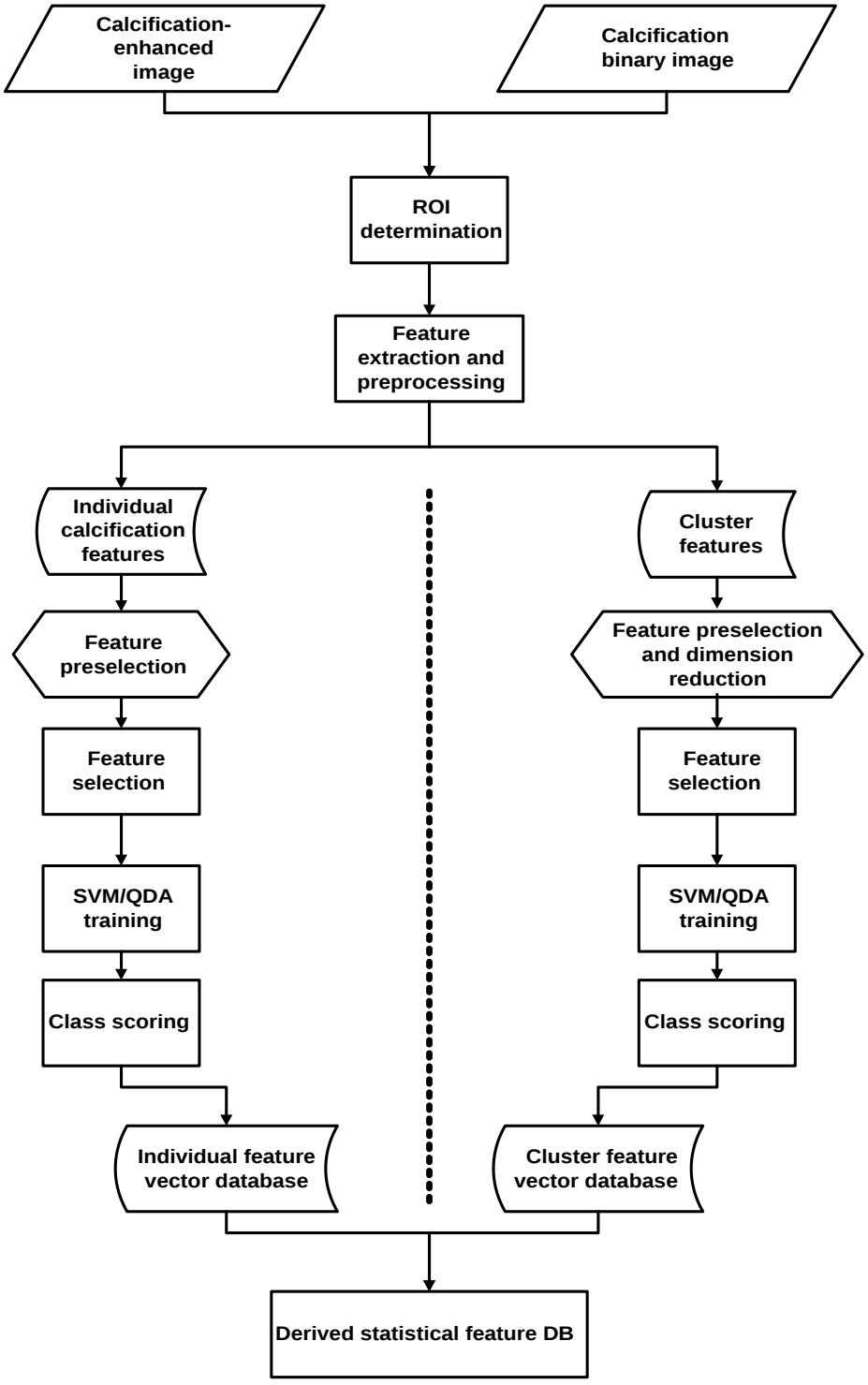
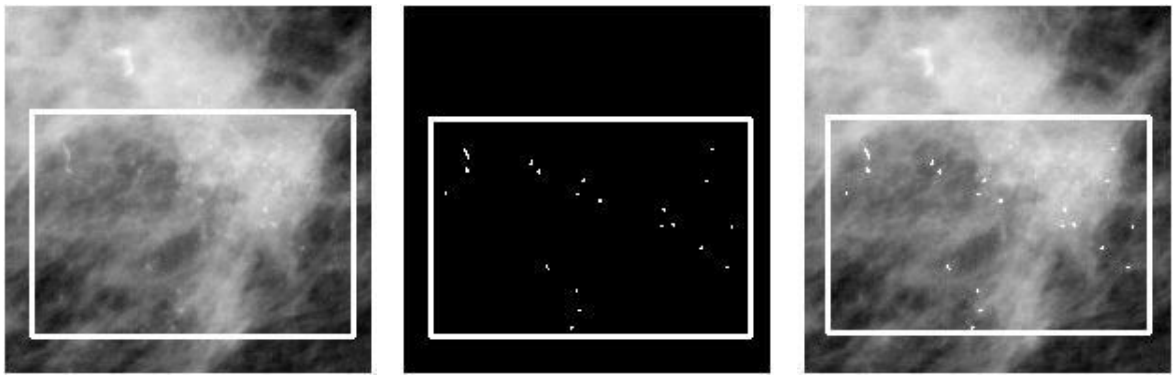


Figure 4.1 Functional diagram of the proposed mammogram image retrieval model. This model takes as input the binary image containing detected microcalcification-like objects as well as the original grey level image that has been enhanced by mean subtraction, as discussed extensively in the previous chapter. The output is the feature vector database

4.3 Region of Interest (ROI) detection

The ROI encompasses the region over which features will be extracted further down the pipeline. This region is drawn around clusters and individual calcification objects. The inputs to the model comprised a segmented binary image depicting calcification objects and its original grey level version as shown in Fig. 4.2. The segmented image is the output of the microcalcification detection phase discussed in the previous chapter. White pixels in the binary image represent calcification objects while the black pixels form the background. A cluster is established where three or more calcification objects exist within an area of $1cm^2$ [51]. If no cluster is found, the calcification objects are considered individually in subsequent processes.

For feature extraction, a bounding box is drawn around individual calcification objects and their containing cluster, where present, on the binary image; the bounding box is padded with pixels from the bordering calcification's boundary pixels. The ROI position and dimensions established from the binary image are superimposed on the gray level image, such that both ROIs have the same coordinates on both versions of the image - these two ROIs will represent the original image in the subsequent steps (See Fig. 4.2). The binary image is simply the segmented version of the grey level image (i.e. containing the detected calcifications).



(a) Original image sample (b) Segmentation results of 4.2(a) (c) Superposition of 4.2(b) on 4.2(a)

Figure 4.2 Microcalcification detection process

4.4 Feature extraction and preprocessing

Feature extraction can be formulated as a mapping from image to feature space, $F : \Re^d \rightarrow \Re^e$, such that a given constraint C is optimized [96]. d and e are the dimensions of the image input space and feature output space respectively, with $d \geq e$ in most applications. The constraint C can be the generalization performance of a classifier, or the retrieval accuracy as is the case in this research.

4.4.1 Feature extraction

In this work, features are extracted to represent individual calcification objects and cluster objects. Information on these two object types is useful in determining the malignancy of a given case as explained in the literature review. The individual calcification feature vector is denoted as $\vec{v}_i$, while the cluster feature set is denoted as $\vec{v}_c$. The choice of features is guided by the need to bridge the semantic gap in the pathological interpretation of microcalcifications between CBIR algorithms and radiologists. In this regard, two features that highly correlate with radiologists' descriptions are extracted as the first set of features. These are: Haralick features, which are extracted from the gray level ROI, and geometric features, which are extracted from the binary ROI [52, 54]. Table 4.1 shows all the feature vector components used to characterize individual microcalcifications, while Table 4.2 lists the features used for microcalcification clusters. These two feature classes have been widely used in calcification characterization and their performance in shape and textural encoding applications is well documented [51].

Haralick features of a gray level image are extracted from the GLCM; they are used for modeling textural characteristics, which are distinctly defined in a calcification-present area of mammograms. The gray level cooccurence matrix $P(i, j|d, \theta)$ encodes the spatial dependencies of tonal intensities i and j, for a given distance d and orientation θ, providing a basis for extraction of second-order statistical features. Given an image I with spatial dimensions $M \times N$ and L grey levels, the GLCM is defined as [97]

$$P(i, j|d, \theta) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f\{I(x,y) = i \text{ and } I(x + d\theta_0, y + d\theta_1) = j\} \tag{4.1}$$

where,

$0 \leq x \leq M-1$, $0 \leq y \leq N-1$, $0 \leq i, j \leq L-1$ and

$$f(v) = \begin{cases} 1, & \text{if v is true} \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

The orientation θ is quantized to four values, which are represented as shown in Eq. (4.3)

$$\theta = \begin{cases} 0^\circ, & \text{if } \theta_0 = 0 \text{ and } \theta_1 = 1; \\ 45^\circ, & \text{if } \theta_0 = -1 \text{ and } \theta_1 = -1; \\ 90^\circ, & \text{if } \theta_0 = 1 \text{ and } \theta_1 = 0; \\ 135^\circ, & \text{if } \theta_0 = 1 \text{ and } \theta_1 = -1; \end{cases} \tag{4.3}$$

This work uses all four orientations shown in Eq. (4.3) and five distances, $d \in [1,3,5,7,9]$. For notational convenience, it is assumed for the rest of the thesis that once the orientation θ and distance d are chosen, $P(i,j|d,\theta)$ will just be represented by $P(i,j)$ and the probability of the cooccurence of the gray levels i and j at distance d and orientation θ will be

$$p_{i,j} = \frac{P(i,j)}{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} G(i,j)} \tag{4.4}$$

The Haralick features used in the proposed model are given in Eqs.

4.5-4.10

$$\text{Maximum probability} = \max(p_{i,j}) \tag{4.5}$$

$$\text{Energy} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{i,j}^2 \tag{4.6}$$

$$\text{Homogeneity} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \frac{p_{i,j}}{1+|i-j|} \tag{4.7}$$

$$\text{Contrast} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{i,j}|i-j|^2 \tag{4.8}$$

$$\text{Correlation} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \frac{p_{i,j}(i-\mu_x)(j-\mu_y)}{\sigma_i\sigma_j} \tag{4.9}$$

$$\text{Entropy} = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{i,j}(-\ln p_{i,j}) \tag{4.10}$$

where,

$$\sigma_i = \sqrt{\sum_{i=1}^{L} p_x(i)(i-\mu_x)^2} \tag{4.11}$$

$$\sigma_j = \sqrt{\sum_{i=1}^{L} p_y(i)(i-\mu_y)^2} \tag{4.12}$$

$$\mu_x = \sum_{i=1}^{L} i p_x(i) \tag{4.13}$$

$$\mu_y = \sum_{i=1}^{L} i p_y(i) \tag{4.14}$$

Geometric features on the other hand describe shape characteristics of clusters or individual calcification objects, which are useful in distinguishing among the various pathologies of calcifications. The five Shape features extracted in this work directly relate to the descriptions used by radiologists to characterize the various calcification properties [51, 84, 98]:

- Area - the total number of foreground pixels

- Compactness - the ratio involving a factor of the object's perimeter and its area. It gives a measure of the roundness of the object
- Orientation - the angle between the x-axis and the major axis of the ellipse of the object
- Eccentricity - the ratio of the distance between the foci of the ellipse and the length of its major. Bigger values imply a higher linearity semblance of the object
- Solidity - refers to the ruggedness of the object, measured as the ratio between its actual area and that of its convex hull

Table 4.1 Individual feature vector $\vec{v}_i$, with dimension $|\vec{v}_i| = 125$

#	**Name**	**Image type**
1	Area	Binary
2	Compactness	Binary
3	Orientation	Binary
4	Eccentricity	Binary
5	Solidity	Binary
6-25	Contrast	Grey level
26-45	Correlation	Grey level
46-65	Energy	Grey level
66-85	Homogeneity	Grey level
86-105	Entropy	Grey level
106-125	Maximum probability	Grey level

Cluster region (CR) in Table 4.2 refers to the grey level ROI identified in the previous Section, the convex hull is drawn around the border calcification objects of the cluster. In the non-clustered ROI, haralick and geometric features are extracted for each individual calcification object.

4.4.2 Feature normalization

At this stage, normalization is applied on both feature sets ($\vec{v}_i$ and $\vec{v}_c$) followed by Principal Component Analysis on the cluster feature vector ($\vec{v}_c$). The $\mathscr{Z}$-score normalization is applied to reduce the undue influence of features having large ranges; it is done using Eq. 4.15 [99],

Table 4.2 Cluster feature vector $\vec{v}_c$, with dimension $|\vec{v}_c| = 141$. The meanings of the abbreviations are as follows: CC - Cluster calcifications, CVH - Cluster convex hull, CR - Cluster region, B - Binary image, GL - Grey level image, μ - mean, σ - Standard deviation.

Feature Description			ROI Object	Image type
#	*Primary*	*Secondary*		
1-2	Area	μ,σ	CC	B
3-4	Compactness	μ,σ	CC	B
5-6	Orientation	μ,σ	CC	B
7-8	Eccentricity	μ,σ	CC	B
9-10	Solidity	μ,σ	CC	B
11	Area		CVH	B
12	Compactness		CVH	B
13	Orientation		CVH	B
14	Eccentricity		CVH	B
15	Solidity		CVH	B
16	Density		CR	B
17-18	Inter-calcification distance	μ,σ	CR	B
19-20	calcification $\rightarrow$ cluster centroid distance	μ,σ	CR	B
21	Number of calcifications		CR	B
22-41	Contrast		CR	GL
42-61	Correlation		CR	GL
62-81	Energy		CR	GL
82-101	Homogeneity		CR	GL
102-121	Entropy		CR	GL
122-141	Maximum probability		CR	GL

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{4.15}$$

where $\tilde{x}$ is the standardized vector, x is the original vector, μ and σ are the sample mean and standard deviation respectively. This process effectively transforms both feature vectors to have zero mean and unit standard deviation.

Given the few clusters that were obtained after feature extraction, the cluster feature vector $\vec{v}_c$ is transformed into a reduced dimension space using PCA, to reduce the dimensions of the resultant vector. PCA seeks a linear combination $Y = \sum_{i=1}^{n} \lambda_i x^{(i)}$ for a column of predictors $x^{(i)}$ of a matrix X such that the columns of Y are linearly independent [100]. The resultant matrix Y is usually ordered with the most significant dimensions first, with this significance defined in terms of variance. By taking the first m significant dimensions of Y, most important information in X can

be retained with the benefit of reduced data.

The dimension of $\vec{v}_c$ was reduced to meet the constraint imposed by the Quadratic Discriminant Analysis (QDA) classifier regarding the minimum number of training samples, as a factor of the feature vector dimension; this constraint ensures effective parametrization during covariance estimation [101]. PCA was not applied to the individual vector $\vec{v}_i$ given that there were enough samples for QDA training. In this work, $\vec{v}_i$ and $\vec{v}_c$ are used as intermediate features, and fed as inputs to the classifiers in Section 4.6.

4.5 Feature selection

The "curse of dimensionality" is a classic issue in CBIR systems, where the performance of such systems expectedly degrades with an increase in the number of features. Indeed, findings have been made that some extraneous features act as noise, degrading the query results of CBIR systems [83, 102, 103]. Ladha and Deepa categorize features into three [102]:

Relevant features have an influence on the output. Their role cannot be overlooked

Irrelevant features have no influence on the output and can be removed without incurring a performance penalty

Redundant features can be substituted by other features

Feature selection plays a three-fold role of (1) reducing the cost of feature extraction, (2) improving classification accuracy and (3) improving the reliability of performance estimate [104, 105]. Feature selection methods are characterized by: a search strategy used to explore the space of hypothesis, a mechanism of proposing feature candidates for the current hypothesis and a measure of evaluating the selected candidate features at any given point. They can be categorized under three classes: filter model, wrapper model and hybrid model. Filter approaches rely on the general characteristics of data to evaluate features based on some discriminating criteria, while wrappers use classifiers and a subset selection approach to measure a feature subset's prediction performance [102]. The Hybrid model is a combination of the two. These three categories have

formed the focus of research in a wide range of applications and datasets [102–108]; a comprehensive discussion on feature selection methods for various biomedical applications can be found in [102, 108], with [105, 107] focusing on microcalcification detection applications.

The optimal feature set was selected from the vectors $\vec{v}_i$ and $\vec{v}_c$ using both filter and wrapper approaches; this combined approach has been successfully applied to the selection of optimal feature subsets for microcalcification characterization [103, 105]. For notational convenience, let us take U_x as the universal set containing all features from class x, and $F(U_x, m)$ as a function that returns m features from the set U_x. The following steps outline the proposed approach for the feature selection process.

1. Convert cluster features to PCA eigen data U_{cPCA}.
2. Normalize by $\mathscr{Z}$-score standardization both individual feature vector U_i and cluster feature set U_{cPCA}.
3. Preselect and rank by decreasing order k features $F_i = F(U_i, k)$ from individual feature set, and l features $F_c = F(U_{cPCA}, l)$ from the cluster set, whose *independent features significance test* result (See Eq. 4.16) $\vec{v}(i) >= 2.0$. This threshold eliminates features that do not have significant discrimination ability.
4. Using the individual and cluster subsets (F_i and F_c) preselected in the previous step, select the optimal feature subsets for the individual and cluster (S_i and S_c respectively) based on prediction performance using Quadratic Discriminant Classifier, using the Forward Selection Feature Search selection strategy.

The preselection step employs a filter approach according to Weiss and Indurkhya called the *independent features significance test* [109]. It involves conducting a hypothesis test on each feature to measure its information value with regards to the separability of the classes. The essence of this step is to remove uninteresting (irrelevant) features with little informative value; this significantly reduces the computational burden for the next step, which is computationally intensive. While this step overlooks dependencies (redundancies) among selected features, it is fast and useful as a preprocessing technique in feature selection applications [104]. Using i to index a particular feature in the feature set $\vec{v}$, the significance of the resultant feature $\vec{v}(i)$ is calculated as follows [109]:

$$Sig(\vec{v}(i)) = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A}{\Sigma_A} + \frac{\sigma_B}{\Sigma_B}}} \tag{4.16}$$

Where μ_A and μ_B is the mean of the features in class A and B respectively, σ_A and σ_B is the standard deviation of feature class A and B respectively.

According to the recommendations by Weiss and Indurkhya [109], all features having significance values less than 2.0 were removed from the final feature subset. They established that features whose significance values according to the test fall below 2.0 are not discriminative enough and can be excluded without incurring a performance penalty. Based on the results of the test, 86 individual features were selected from the original set $\vec{v}_i$ for scoring above 2.0 (see Fig. 4.3). As mentioned in the previous section, $\vec{v}_c$ is transformed using PCA before significance testing is done (Eq. 4.16). The significance test is thus performed on the PCA coefficients. As seen in Fig. 4.4, 18 PCA coefficients score above the test cutoff mark of 2.0.

The second selection step applies a wrapper approach to remove redundant features using the Quadratic Discriminant Analysis (QDA) classifier and the Forward Selection Forward Search (FSFS) method. The QDA classifier is employed at this stage because of its relatively inexpensive time cost, as contrasted to the SVM classifier. The FSFS search strategy incrementally adds features to an initial null set until further addition cannot minimize the error rate [102]. The results of the feature selection process are shown in Fig. 4.5 and 4.6.

4.6 Classifier training

The SVM and QDA models were trained using the selected feature sets. Three parameters needed to be established for the SVM classifier: the kernel type, its associated parameters and the constraint value C. The linear, polynomial and rbf kernels were selected for their high recommendation in similar works [110]. The final kernel chosen from the three was that which gave the minimal classification error. This study used an unconstrained linear optimization method to establish the optimal parameter values. Initially, a search was conducted through a set of equally spaced linear values, followed by fine-tuning of the selected parameters by searching random values around them. Algorithm 3 shows the steps followed in fine-tuning the parameters of the SVM classifier.

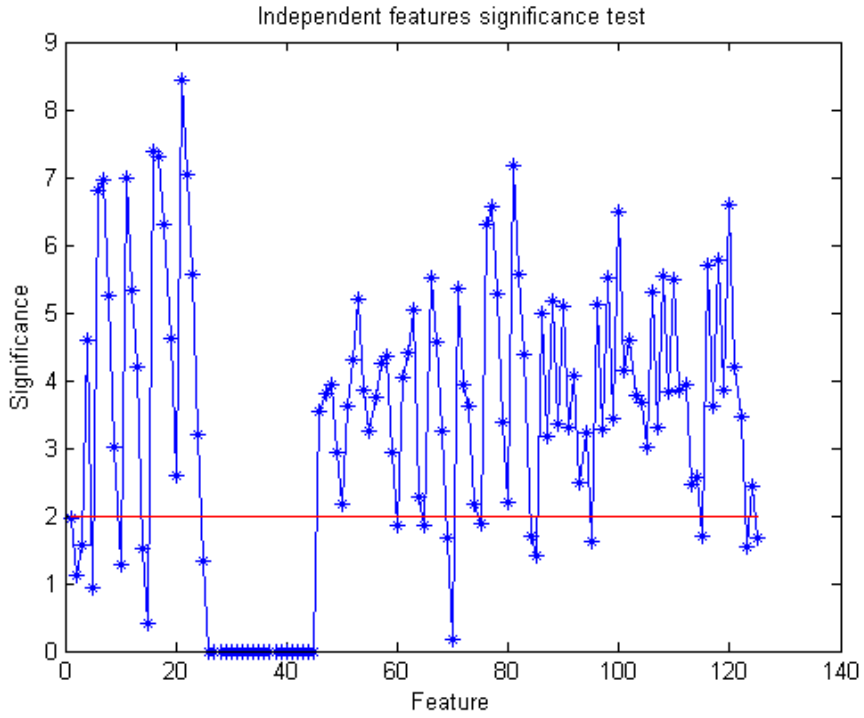


Figure 4.3 Results obtained from applying the independence significance test on the individual feature vector. The horizontal line specifies the threshold for significant discrimination ability. Values below the line imply that the feature in question cannot discriminate between the two classes. The higher the value, the more a given feature can differentiate between malignant and benign samples.

Algorithm 3 The steps involved in the training of the SVM classifier

function TRAINSVM(Data)
    [TrainData,TestData] = *partitionData*(Data,10) ▷ randomly partition Data into 10 sets with equal class representation
    Cost $\leftarrow$ *RandomGenerator*(20), Scale $\leftarrow$ *RandomGenerator*(20) ▷ Initialize Cost and Scale parameters to 20 random real values
    for all Kernel $\in \{'linear', 'polynomial', 'rbf'\}$ **do**
        for $i = 1 \rightarrow 20$ **do**
            SVMModel = *trainModel*(Cost[i], Scale(i), Kernel,TrainData)
            Error = *getClassificationError*(SVMModel,TestData)
            CE[i] $\leftarrow$ *Average*(Error)
        end for
        CE_{min} = *getIndex*(*Argmin*(CE)) ▷ Get index of minimum classification error
        $Cost_{opt}$ = Cost[CE_{min}], $Scale_{opt}$ = Scale[CE_{min}]
        SVMModel = *trainModel*($Cost_{opt}$,$Scale_{opt}$,Kernel,Data) ▷ train using optimum parameters on all data
    end for
    return SVMModel ▷ return optimal model
end function

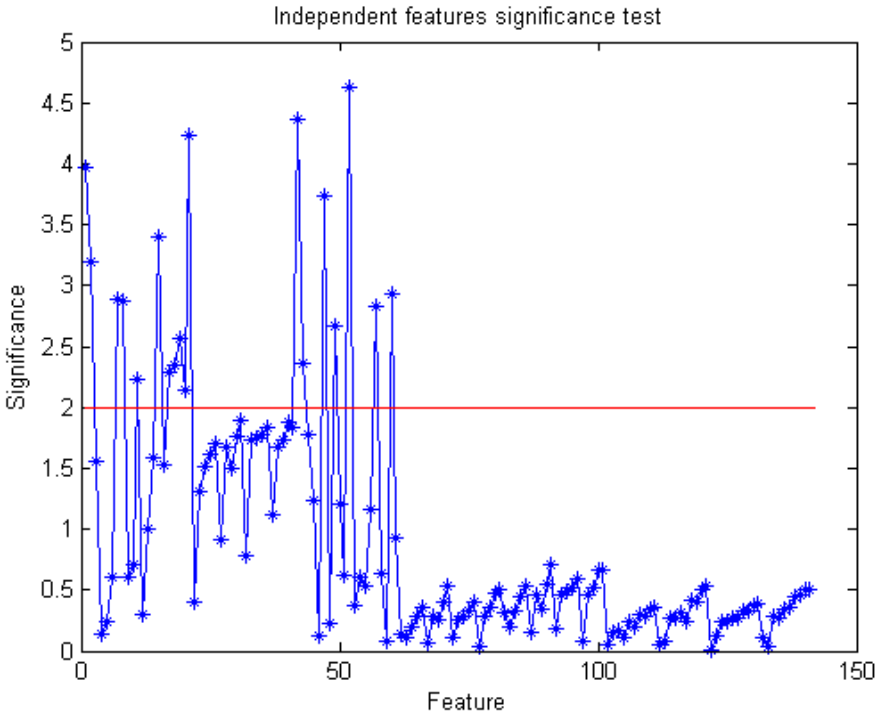


Figure 4.4 Results obtained from applying the independence significance test on the cluster feature vector. The horizontal line specifies the threshold for significant discrimination ability. Values below the line imply that the feature in question cannot discriminate between the two classes.

The database is divided into ten folds for the training and testing datasets. In accordance with the fundamentals of classification problems, the trained model's performance is tested on samples that were not included during the training, so as to ensure a better generalization performance. The motive of the SVM training is to establish the best kernel type as well as the cost and scale parameters of the appropriate kernel. An array of 20 parameter values are chosen for the cost and scale parameters; these are used to train the SVM model at the end of which the best performing parameter pair is selected. For the initial coarse parameter search, the bounds for the range of values considered in the parameter search for σ and C were taken from related work [110]. Specifically, the parameter search for the parameters considered the following values: $C \in \{2^{-i} | i = -3, -2, ..., 15\}$ and $\sigma \in \{2^{-j} | j = -12, -2, ..., 4\}$. The parameter pair that returns the minimal classification error is chosen as the best and subsequently taken as the representative for that round. This process is repeated for each of the ten folds of training and test samples for more robustness of the model.

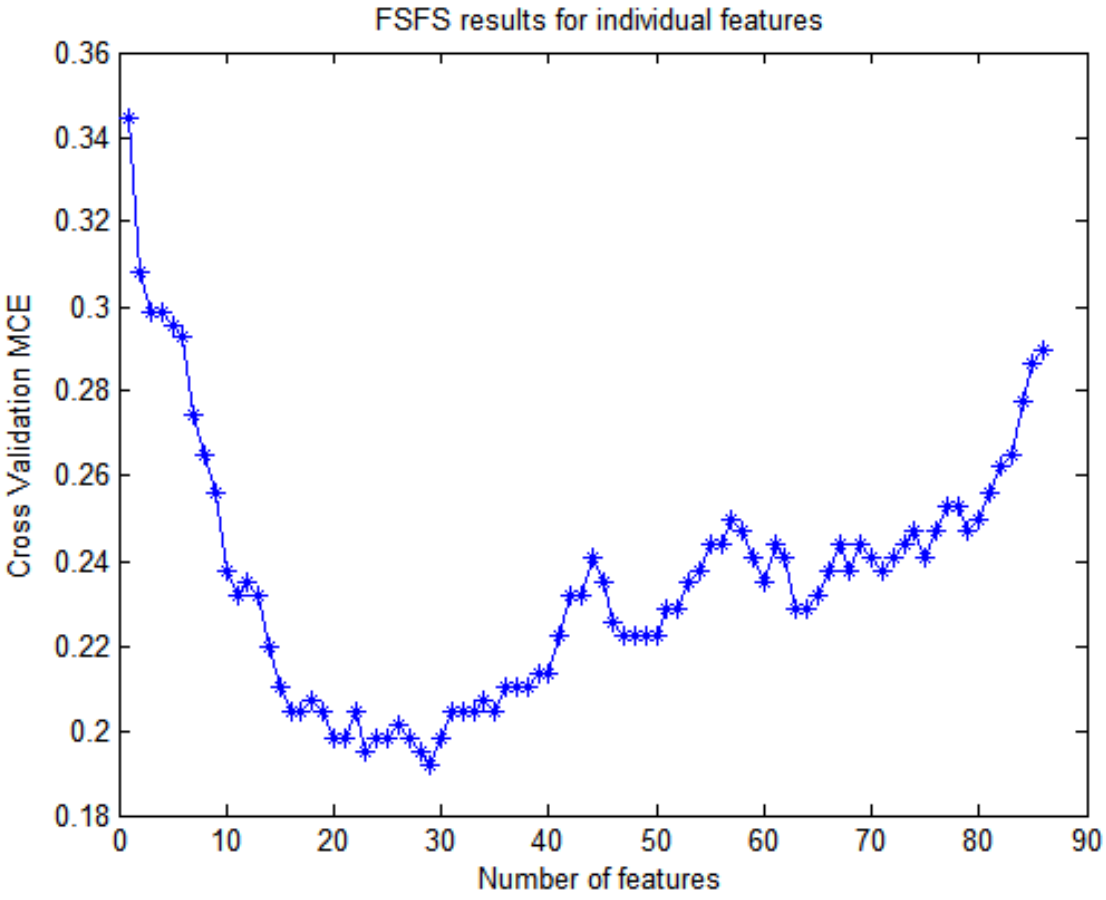


Figure 4.5 Feature selection results for individual calcification features ($\vec{v}_i$) based on the Forward Selection Feature Search method (FSFS). Lower values of the cross validation classification error (y-axis) show better performance of the selected subset. The global minimum forms the cutoff for the best subset and can be seen at the 29 mark on the x-axis.

The optimal parameter pair (in terms of reducing the classification error) is finally adopted as the parameters of the SVM model. The optimum model was established to be the polynomial kernel with the following parameters: $\sigma = 2.7803$ and $C = 1000$.

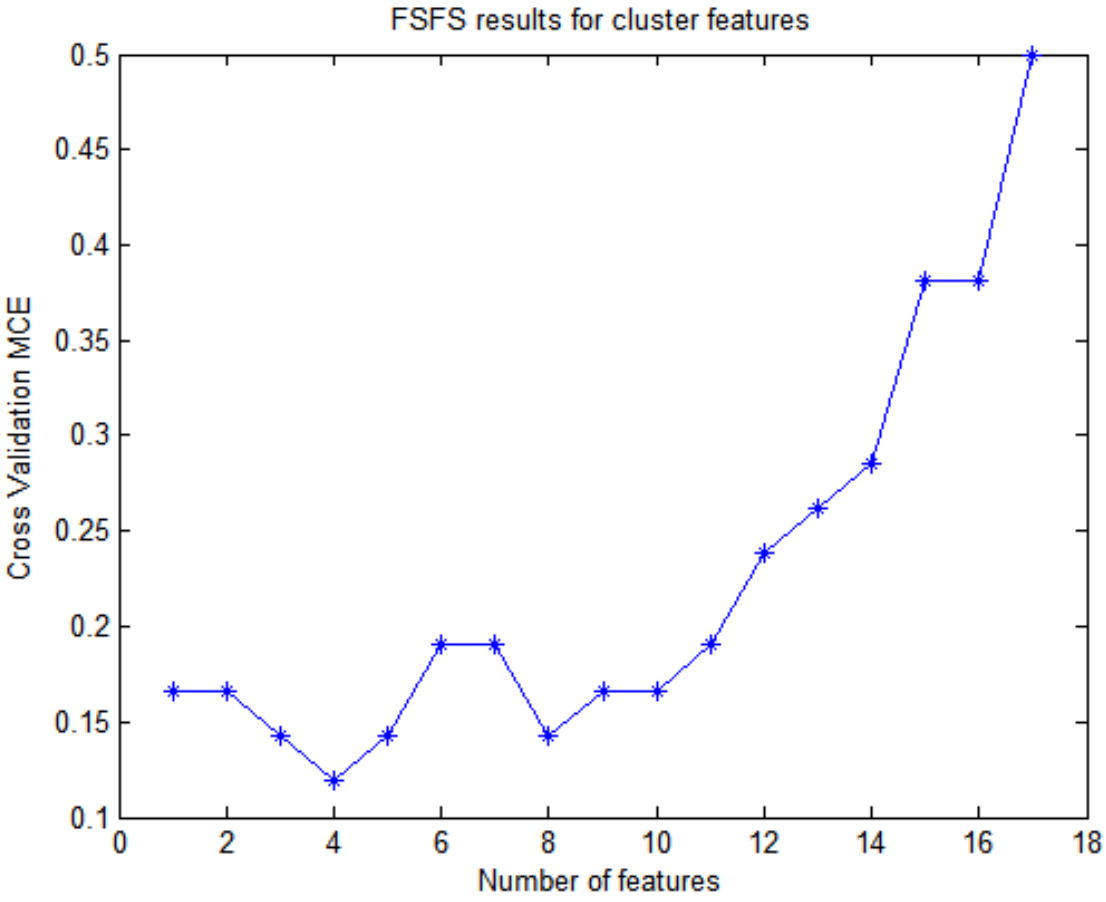


Figure 4.6 Feature selection results for cluster features ($\vec{v}_c$) based on the Forward Selection Feature Search method (FSFS). The best classification error according to the graph is at around 11.9%, which is attained with a feature dimension of 4.

4.7 Classifier scoring

The final feature vector is derived from classifier scores, unlike most related research works that directly employ primary features (Haralick, Wavelet, Geometric, etc) as input to the k-NN algorithm [52, 54, 69]. The trained SVM and QDA classifiers are used to generate scores which are used for creating the final feature vector in this stage. While classifier scores have been used

before in the literature, this study extends the notion to include statistics on the scores as well. Furthermore, the contribution of a given feature is weighted based on its significance test described in Eq. (4.16). Scores for both the cluster region and the individual calcifications are considered. Table 4.3 presents the features used as the final vector set, as well as their significance and relative discrimination strength.

Table 4.3 The final feature vector is a combination of classifier scores as well as basic first order statistics on them. This table also shows the individual performance of the features based on the *independent features significance test* described in Section 4.5. Feature relevance in the second column describes the total contribution of that feature, in terms of discrimination ability, to the entire feature set.

Feature	Feature relevance	Description
1. SVM score	3.75(0.05%)	Cluster's SVM score
2. μ_{SVM}	17.82(0.25%)	Average of SVM scores for all ROI calcifications
3. μ_{QDA}	16.16(0.23%)	Average of QDA scores for all ROI calcifications
4. μ_{SVM+}	12.81(0.18%)	Average of SVM scores for positive ROI calcifications
5. μ_{QDA+}	8.90(0.13%)	Average of QDA scores for positive ROI calcifications
6. σ_{QDA}	3.46(0.05%)	Standard deviation of QDA scores for all ROI calcifications
7. σ_{SVM+}	3.85(0.05%)	Standard deviation of SVM scores for positive ROI calcifications
8. σ_{QDA+}	4.04(0.06%)	Standard deviation of QDA scores for positive ROI calcifications

"Positive ROI calcifications" as mentioned in Table 4.3 refers to those calcifications that are classified as positive by the classifiers. A calcification is considered positive only if its score for both the QDA and SVM classifiers is greater than 50%.

4.8 Similarity measurement and ranking

The voting k-Nearest Neighbor (k-NN) classifier is finally used to assign a class to the query image based on the ranked results. The k-NN classifier is a non-parametric classification technique that assigns to a sample the class represented by a majority of its k neighbors [105]. This method assumes all instances in the database as points in an $n-$dimensional space and calculates the distance d between the query vector q and all the other samples, returning the set of k vectors in increasing

order of distance [111]. The distance d is commonly referred to as the measure of dissimilarity and is defined as a mapping $d : X \times X \rightarrow \mathbb{R}^+$.

The similarity (or dissimilarity) metric is a critical factor that can affect retrieval precision [66, 101] in biomedical applications. Similarity metrics based on feature vector distance measurements rank retrieval results based on the images' feature vectors distance. Similarity measures $\mathfrak{d}$ based on distance measurements must meet the following requirements,

- *Positivity:* $\mathfrak{d}_{\mathrm{i,j}} > 0$
- *Symmetry:* $\mathfrak{d}_{\mathrm{i,j}} = \mathfrak{d}_{\mathrm{j,i}}$
- *Identity:* $\mathfrak{d}_{\mathrm{i,i}} = 0$

where i and j are the feature vectors. A commonly used family of metrics is the Minkowski distance $\mathscr{L}_p$, which is defined as,

$$\mathfrak{d}_{r,s} = \left(\sum_i |x_{ri} - x_{si}|^p \right) \tag{4.17}$$

where x_{ri} is the i^{th} component of the feature vector of r. This work uses the Euclidean distance for calculating the dissimilarity between database samples. For instance, if we consider a query image represented by its feature vector, $\mathbf{q} \in R^d$, of dimension d, such that, $\mathbf{q} = [q_1, q_2, \ldots, q_d]^T$, then the Euclidean distance $\mathscr{L}_2$ between the query vector and a particular database image with the feature vector $\mathbf{f}$ is defined as follows [17]:

$$\mathscr{L}_2 = \|\mathbf{f} - \mathbf{q}\|^2 = \sum_{i=1}^{d} (f_i - q_i)^2 \tag{4.18}$$

In the case where more than one database samples have the same distance d to the query, the algorithm returns the first k images of the result as they are ordered in the database. The k value of the k-NN classifier was taken from the set $\{1, 3, 5, 7, 9, 11\}$.

4.9 Conclusion

Content-based image retrieval is a potentially useful technique in supporting diagnostic decisions by availing similar cases with known pathology to radiologists. Accuracy and efficacy are required for CBIR systems to be adopted in regular medical practice. This chapter presented an improved model for the retrieval of mammograms based on their pathology. The highlight of this chapter is the combination of classifier scores, with statistics on the same to construct an effective feature vector for improving the accuracy of the retrieval of mammogram cases based on pathology and more specifically, the malignancy of clusters, where present. The feature characterization model was further improved by appropriate weighting of the features based on results of their individual discriminatory ability.

Chapter 5

Parallel extraction of features

Time complexity is a significant issue especially with algorithms that run on large data, or are intrinsically computationally expensive. Parallel computing is a probable solution to this kind of challenge. Rather than simple application of techniques, a parallel model has to be efficienctly designed to take advantage of available concurrency of the problem's subtasks while keeping overheads at a minimum. This chapter presents a model that exploits data-parallelism with dynamic task mapping towards reducing the time complexity for feature extraction using the Message Passing Interface (MPI) framework. The proposed model factors in the latency cost in the design of a task mapping strategy that ensures a sustainable, near optimal performance.

5.1 Feature extraction

The task of this parallel model, as shown in Fig. 5.1, considers three feature extraction methods considered in the previous chapters. One of the aims in the design of feature extraction methods is reducing the computational complexity of the process [82]. In machine learning applications, the accuracy of a feature extraction model is measured by its ability to reduce the generalization error of a given classifier, which is a factor of its inter-class variance [18, 112]. Computational complexity constraints are equally important in ensuring the extraction process is efficient in terms of resource utilization.

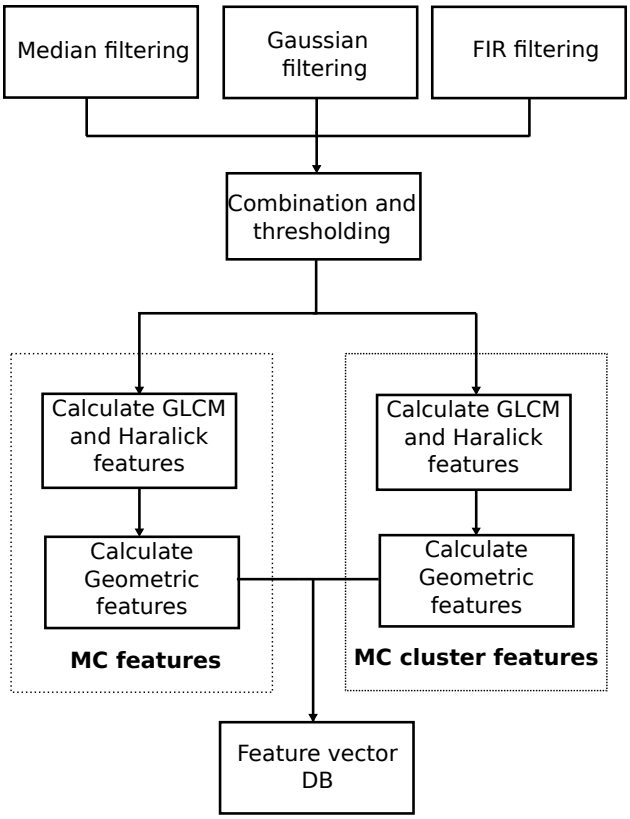


Figure 5.1 Feature extraction processes.

5.2 Parallel model

This work considers a parallel computation model based on data-parallelism with dynamic task mapping and the conventional First Come First Served (FCFS) scheduling policy. The task partitioning method is designed to reduce data dependencies in this manner: the entire database is considered as the input space N, with the data decomposition process partitioning the input space into complete images $n_i \in N | i = 1, ..., |N|$, which form the input data for the processes. Fig. 5.2 shows the task-dependency graph of the proposed model.

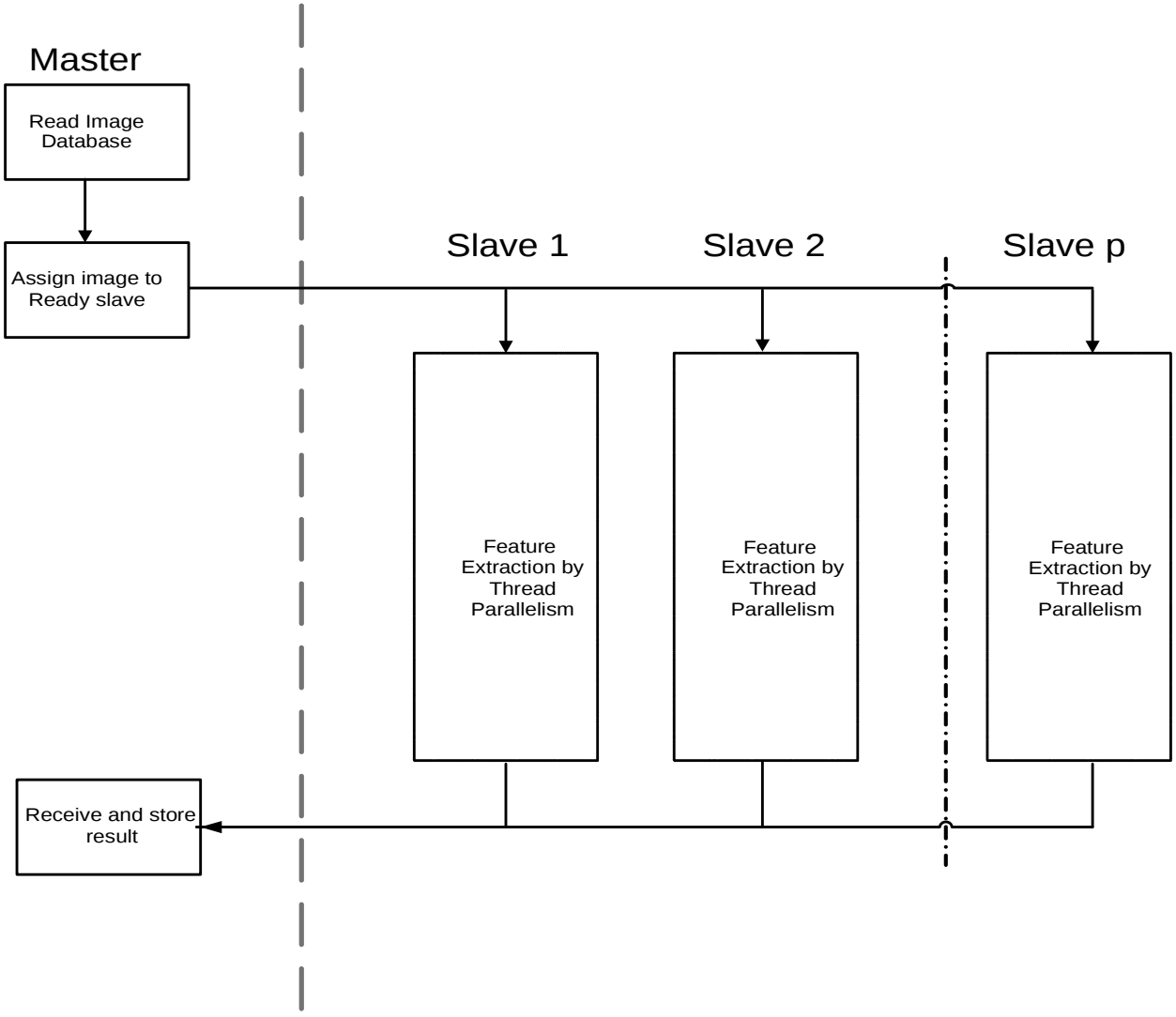


Figure 5.2 Task-dependency graph of the global master parallel computation model.

The decomposition scheme employed results in limited data dependencies (See Fig. 5.2) among the tasks making this a coarse-grained problem, effectively maximizing the degree of concurrency of the tasks. A coarse granularity can be counter-productive if the tasks are computationally expensive relative to the communication cost; it is thus imperative to optimize the task load at each compute node. It can easily be established that the maximal degree of concurrency for our scheme is equal to the database size. It can be posited that a larger database size will lead to a better relative performance as the execution time per task is amortized by the increasing concurrency of tasks; however, we expect the increase in performance as a factor of increasing node (or process) size to be constrained by communication and synchronization costs among the Master and Slave

nodes.

The database considered in this work contains images of a standard size of dimension 1024×1024 pixels, this implies an evenly balanced task load for all nodes, on the assumption that they have similar processing power. Our scheme employs centralized mapping of tasks using a Master-Slave approach as shown in Algorithm 4. In this scheme, one process is designated as the master, which is responsible for partitioning (decomposing) the input data space and allocating the same to compute nodes that signify their willingness to receive tasks. Worker nodes are assigned tasks on a first-come-first-served basis, which means that nodes with superior processing power will get more work assigned given that they will make more requests upon task completion. This approach better adapts the model to an inhomogenous environment where the processing power differs among the compute nodes.

Algorithm 4 MPI-based centralized task mapping for load balancing. Slaves are assigned tasks on a first-come-first-served basis

```
1:  if MASTER then
2:      DB ← ReadImageDatabase()                    ▷ Read in the image database
3:      for all Image ∈ DB do
4:          SendImageToReadyProcess(Image)
5:          ResultArray = FetchResultFromReadyProcess()
6:      end for
7:  end if
8:  if SLAVE then
9:      while MasterHasTasks() do
10:         Data = GetImageFromMaster()
11:         ExtractFeatures(Data)
12:         SendFeatureVectorToMaster()
13:     end while
14: end if
```

Priority is given to the assignment of tasks. This means that while there still exist pending tasks and ready compute nodes, then the operation conducted the master node is task assignment. The priority towards task assignment is true even in the scenario of multiple master threads, as all are dedicated towards serving tasks to ready nodes. This is to reduce the time spent waiting for tasks by ready compute nodes. Compute nodes that have completed computation immediately send the results back to the master node in a blocking call. This means they are suspended until the master is ready to receive their results.

5.3 Optimizing latency

The communication overhead c_o, incurred during the transmission of an n-byte message in a parallel model is defined as a function of the network latency, γ, and the transmission speed β, as [78],

$$C_c(n) = \gamma + \beta n \tag{5.1}$$

C_c can be a significant contributing factor to the idling of compute nodes [78]. This is typical in non-parallel I/O systems where access to the input data resource is restricted to one node in a cluster. Assuming that communication can take place only between two nodes at any given time (point-to-point), the idle time can be significant for nodes down the communication pipeline as they wait to be served. This waiting time can be considered as a form of latency, considering that it is the time taken before the node receives its task from the master. The "waiting latency" is referred to in this work as "model latency", signifying the latency introduced by the design of the cluster communication model, in contrast to the conventional definition that considers the underlying hardware system. We consider it latency due to the model because it is dependent on the task size as well as mapping scheme of the model rather than external factors. With the aid of Fig. 5.3, the problem is described in the following paragraphs.

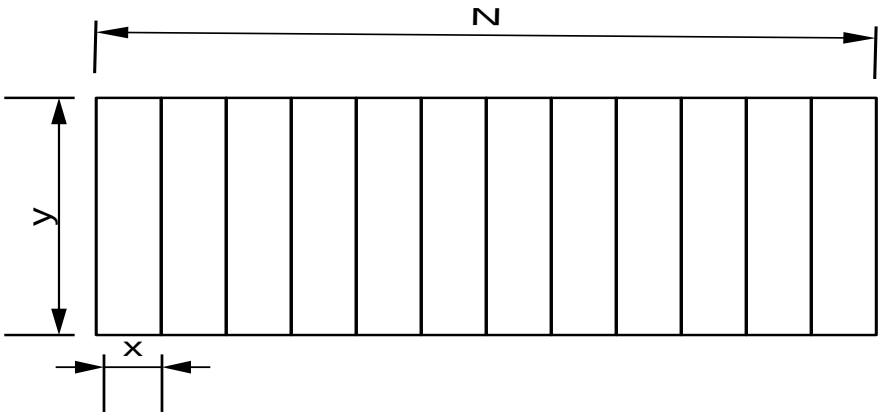


Figure 5.3 Coarse partition scheme considering database as input space

At the distributed computing model level, this work considers data parallelism where each node executes the same instructions on a different data segment. This work considers a database of

images D as the input space, with a coarse partition scheme on the database that simply gives back its constituent images $D = \cup_{i=1}^{N} D_i$. As illustrated in Fig. 5.3, all images have the same dimension $n_i = |I| = x \times y \mid i = 1, ..., N$. N is the number of partition blocks.

A non-parallel I/O computation on the dataset would necessitate N communications which individually incur the communication overhead C_c defined in Eq. 5.1, giving the total overhead C_{ct} as

$$\begin{aligned} C_{ct} &= (\gamma + \beta n_1) + (\gamma + \beta n_2) + \ldots + (\gamma + \beta n_N) \\ &= \sum_{i=1}^{N} (\gamma + \beta n_i) \end{aligned} \tag{5.2}$$

given that γ and β are constants, we can reduce the above equation to

$$C_{ct} = N\gamma + \beta(x \times y)N \tag{5.3}$$

It can easily be seen that the penalty due to model latency increases linearly with an increase in N and vice versa. Let $0 < k \leq N$ be an integer, $\tilde{N} = N/k$ the new number of blocks, $T = x \times y$ and $\tilde{T} = kT$ the new block size; this gives the new overhead function as

$$\begin{aligned} C_{ct}(k) &= \frac{N}{k}\gamma + \beta\frac{N}{k}kT \\ &= \tilde{N}\gamma + \beta\tilde{N}\tilde{T} \end{aligned} \tag{5.4}$$

It has to be noted that the operation N/k is integer division - for simplicity, subsequent discussion assumes that k divides N without a remainder, in which case we can rewrite Eq. 5.4 as

$$C_{ct}(k) = \frac{N}{k}\gamma + \beta NT \tag{5.5}$$

While the bandwidth cannot be varied, Eq. 5.5 nonetheless enables the reduction of model latency penalty, which is the first term of the equation, by choosing a larger k. Indeed, by setting $k = N$ the latency is reduced to the same value as sending a single block in Fig. 5.3. Eq. 5.5 can be seen as an optimization problem involving the minimization of C_{ct} over the domain of k. A naive solution would be set $k = N$ which would imply sending all data I to a single compute node; however, that would defeat the purpose of data parallelism given that only one worker node is involved in the computation.

Therefore, choosing the k value in Eq. 5.5 has to be balanced with the need to sustain fine-grained decomposition to allow the maximum speedup possible. A smaller k value ensures a finer decomposition, which increases the degree of concurrency, and vice versa. Our objective thus is to minimize the model latency cost without significantly affecting the degree of concurrency. Assuming that $I = x \times y$ is the finest decomposition block size possible in Fig. 5.3, then N would be the maximal degree of concurrency in the original partition scheme, and $\tilde{N} = N/k$ in the new partition scheme. Setting $k = 1$ would have no effect on the concurrency, which is reduced by a factor of N/k. Suppose it takes $t(I_o) = t_o$ seconds to process a single block of the original scheme I_o, then it would take $t(I_k) = kt_o$ for the new block I_k, after repartitioning the input space ($\tilde{N} = N/k$) according to Eq. 5.5. The overall time for the whole process using p compute nodes, ignoring the overhead costs, could be calculated according as

$$t(N,k,p) = \frac{N}{kp}kt_o = \frac{N}{p}t_o \tag{5.6}$$

There is a practicality constraint on p defined as $0 < p \leq \tilde{N}$. This stems from the fact that at this decomposition level, I_k is atomic and can only be assigned in its entirety to a single compute node, therefore the number of assignable compute nodes p can only be as much or less than the available tasks $\tilde{N}$, with extraneous nodes deemed irrelevant. Thus, k has the effect of reducing the number of assignable processors. The assumption on Eq. 5.6 is that $N \bmod kp = 0$; if this is true, the overhead cost due to latency can be reduced. If the constraint does not hold, then the time cost of the partitioning scheme would be calculated as per Eq. 5.7, which is the general form for any value of k (including the default allocation scheme $k = 1$).

$$t(N,k,p) = \left\lceil \frac{N}{kp} \right\rceil kt_0 \tag{5.7}$$

The communication overhead c_o in Eq. 5.1 is just one of the overhead costs in a parallel model. The overall overhead cost in the system can be estimated as

$$C_t = pT_p - T_s \tag{5.8}$$

or according to our model

$$C_t = \frac{N}{p} t_o - T_s \tag{5.9}$$

The overall time can thus be estimated as

$$t(N,k,p) = \frac{N}{p} t_o + C_t \tag{5.10}$$

since we know the communication overhead C_{ct} to be part of the overall overhead C_t, we can rewrite the Eq. 5.10 to be more specific as

$$t(N,k,p) = \frac{N}{p} t_o + C_{ct} + (C_t - C_{ct}) \tag{5.11}$$

with $(C_t - C_{ct})$ representing other overhead costs. We then calculate the overall running time as

$$t(N,k,p) = \frac{N}{p}t_o + \frac{N}{k}\gamma + \beta\frac{N}{k}kI + (C_t - C_{ct}) \tag{5.12}$$

Given that $(C_t - C_{ct})$ has a small cost compared to the C_{ct}, we can ignore it. The running time $t(N,k,p)$ gives us our final objective function, which we seek to optimize by finding the value of k that minimizes the function value as

$$Argmin_k(t(N,k,p)) = \frac{N}{p}t_o + \frac{N}{k}\gamma + \beta NT \tag{5.13}$$

Subject to,

$$0 < p \leq \frac{N}{k} \tag{5.14}$$

$$0 < k \leq N \tag{5.15}$$

The solution to the minimization of $t(N,k,p)$ with respect to k is

$$k = -\sqrt{N\gamma} \tag{5.16}$$

Eq. 5.13 can be further simplified by assigning the maximum compute nodes to the task such that $p = N/k$. This is reasonable in the sense that increasing the nodes will reduce the processing time, which is the first term. This also allows us to get rid of the first constraint, giving us the optimization problem defined as

$$Argmin_k(t(N,k,p)) = kt_o + \frac{N}{k}\gamma + \beta NT \tag{5.17}$$

$$0 < k \leq N$$

The value of k_s that satisfies Eq. 5.17 based on the derivative $\delta t/\delta k$ and the constraint $k > 0$ is chosen as

$$k_s = \sqrt{\frac{N\gamma}{t_o}} \tag{5.18}$$

Increasing the task size to k_s as given in Eq. 5.18 should reduce the overall time cost provided the constraint $N \bmod k_s p = 0$ is true. This is true because the load allocation to all processors is equal. When the constraint is violated, the partitioning of the task set by k becomes counter-productive. Before the partitioning scheme based on k is implemented on the pending task set N, a check is done to ensure no performance penalty due to load imbalance.

$$k = \begin{cases} k_s, & \left\lceil \frac{N}{kp} \right\rceil k t_0 \leq \frac{N}{p} t_0 \\ 1, & \text{otherwise} \end{cases} \tag{5.19}$$

5.4 Conclusion

The benefit of a parallel model can be undone by overheads intrinsic to the coordination of multi-computer resources being leveraged to solving a particular task. These overheads impose a ceiling on the maximum speedup achievable by a parallel processing model. The reduction of these overheads is a paramount design goal by researchers in increasing the efficiency of parallel models. The communication overhead is one such overhead; it describes the penalty implied in the transmission of messages among compute nodes. This chapter discussed a proposed methodology that aims to reduce the communication overhead by optimizing task assignment. Specifically, the assignment of task chunks to slave nodes by the master node considers historical data about the bandwidth and latency. These variables form part of a minimization problem whose solution should mitigate the communciation cost overhead, ideally increasing the efficacy of the parallel model. One image is considered as a task chunk in a coarse partition approach, with the chunks being distributed to the compute nodes using the message passing interface (MPI). The parallel task is extraction of features described in Chapter 4. At the node level, threads can be spawned to extract features using OpenMP thread parallelism. The results of this methodology as well as the detection of microcalcifications and feature extraction are discussed in the next chapter.

Chapter 6

Results and Discussion

This chapter presents and discusses the results of the proposed system with regards to microcalcification detection (Chapter 3), feature extraction (Chapter 4) and parallel computing (Chapter 5). All the images used in this work were sourced from the MIAS database [92], which has been widely used by related research [6, 42, 56, 59, 113, 114]. The MIAS database comprises of a total of 161 pairs (giving a total of 322 image cases) of film of mixed pathology that were selected from the United Kingdom national breast screening program. The images were digitized to a spatial resolution of $50\mu m$ with a Joyce-Loebl microdensitometer having a linear response optical density range of between 0.0 and 3.2. The images were taken in Medio-Lateral Oblique view. The images have been categorized into breast type and film category and come with a ground truth established by medical experts. The MIAS database has 29 ROIs containing microcalcifications, of which 15 are Malignant. These ROIs are spread over 25 different cases.

6.1 Performance Metric

Multiple metrics were used in a complementary fashion to give a wider assessment of the proposed model; this circumvents the incompleteness of singular metrics as test validity descriptors [115] and facilitates wide comparison with related work. Specifically, the performance of the models was benchmarked using Sensitivity (or True Positive Rate), Specificity (is equal to 1-False Positive Rate

(FPR)), Accuracy and Positive Predictive Value (PPV). These metrics were chosen because of their wide use in related applications [105, 110] and the valuable information on system performance that they capture [115, 116].

The metrics used for the evaluation of the model are based on the 2×2 confusion matrix of the concepts of true/false positives/negatives, as shown in Table 6.1 [117, 118]. The concenpts of True/False positive/negative relate to how the decision of the algorithm coincides with the true clinical decision. Specifically, True Positive (TP) is the number of correct classification of a given mammogram as positive. True Negative (TN) is the correct classification of a given mammogram as negative. False Positive (FP) is incorrectly classifying a negative mammogram as positive and False Negative (FN) is incorrectly classifying a positive mammogram as negative. Positive in this context means that a given mammogram has microcalcification clusters as determined by a radiologist. Our measurement metric is based on Karssemeijer's criteria for counting true and false positives [119] as follows: a true cluster is flagged if three or more objects are detected within a radius of 1cm; a False Positive (FP) cluster is counted if none of the objects found in the cluster are inside the truth circle. The truth circle is determined from the ground truth provided together with the MIAS database.

Table 6.1 2×2 confusion matrix depicting True Positives (TP), False Positives(FP), True Negatives (TN) and False Negatives(FN)

		Disease	
		Positive	**Negative**
Test	**Positive**	TP	FP
	Negative	FN	TN

Having established the values of the contingency table, sensitivity, specificity, accuracy and the positive predictive value (PPV) are calculated as in Eq. 6.1-6.4,

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{6.1}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{6.2}$$

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \tag{6.3}$$

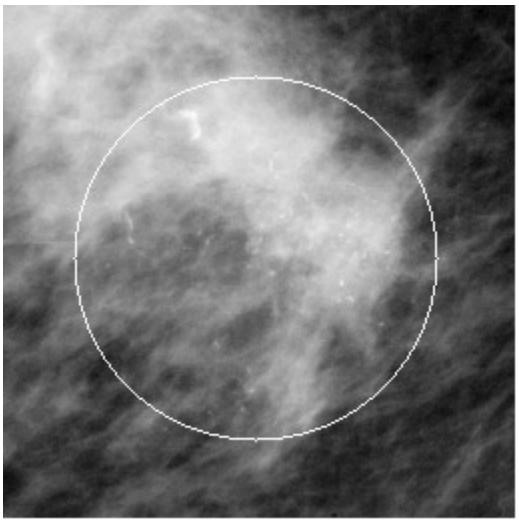
$$\text{PPV} = \frac{TP}{TP+FP} \tag{6.4}$$

Sensitivity (also called the True Positive Rate/Fraction i.e. TPR/TPF) gives a measure of the probability that the algorithm will correctly classify an unseen positive query, while specificity is the probability that it will correctly classify an unseen negative query [115]. A high sensitivity model implies that it is unlikely to miss a test, and is usually preferred in screening for a disease [115]. A high specificity value is equally desirable as it implies a lower probability of false positives. The Positive Predictive Value gives the probability of the sample being truly positive, by considering the prevalence of the disease; it is useful since a positive classification score does not automatically imply the presence of disease, but rather varies with the prevalence of the disease within the population sampled. For instance, a highly sensitive test will have many FPs if the disease prevalence is low [120]. A good model should score high values in all the aforementioned metrics.

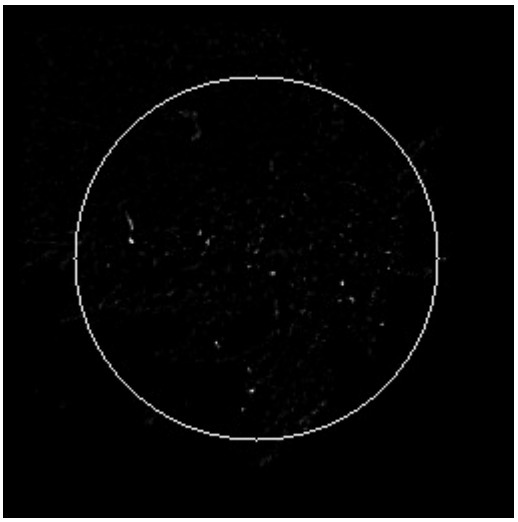
6.2 Microcalcification detection

The first phase is the detection of microcalcifications discussed in Chapter 3. For validation of this model, the experimental setup considered 27 ROIs classified malignant due to microcalcifications and 99 (randomly chosen) classified as normal, from the MIAS database with each ROI having a resolution of 256×256 pixels; the cluster containing the abnormality is centered on each ROI. The 27 ROIs were chosen from the database because they are clearly described in the ground truth with regards to center of abnormality and cluster size; this information lacks in the 2 ROIs that were left out.

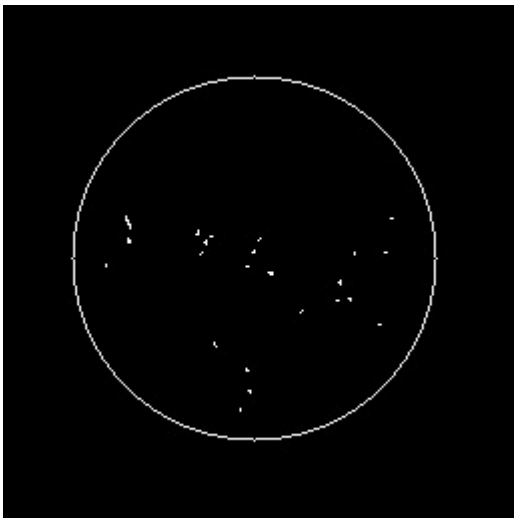
Considering the visual results are presented in this section, the white circle in the Fig. 6.1-6.3 is an overlay delineating the cluster boundary as traced by an radiologist expert, according to the accompanying ground truth. Fig. 6.1-6.3 illustrate three of the scenarios presented in the results. In Fig. 6.1, all the microcalcifications in the clusters in the truth circle have been detected. The type of the breast tissue is Fatty-Glandular. The Benign calcification has been detected in the Fatty-Glandular ROI *mdb218* in Fig. 6.2. In Fig. 6.3, all the microcalcifications in the cluster have been detected. However, the proposed model also marked out other objects (lower right of the image) as calcifications.



(a) Original image

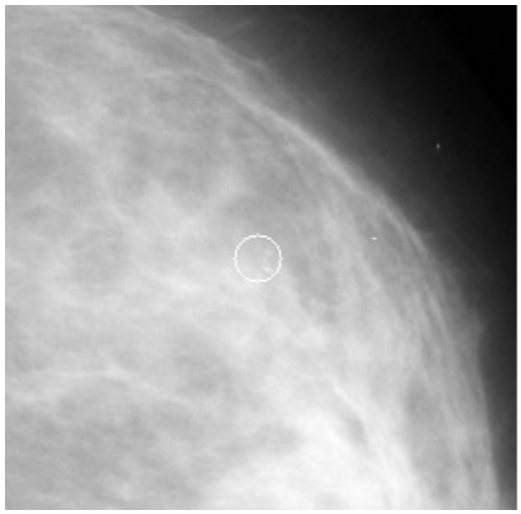


(b) Enhanced image after linearly combining the output of wavelet analysis, Median and Gaussian filtering

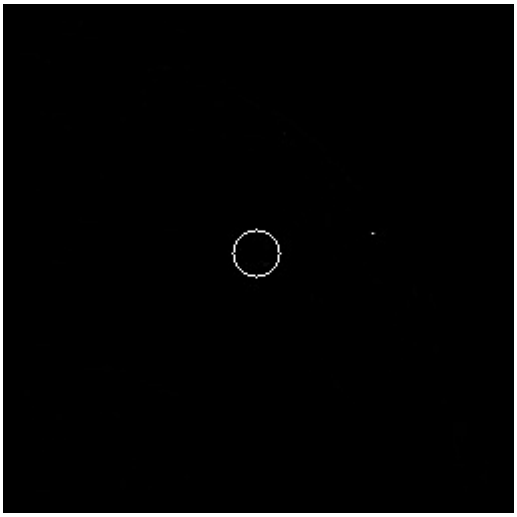


(c) represents the final image, after thresholding and applying all the post-processing operations

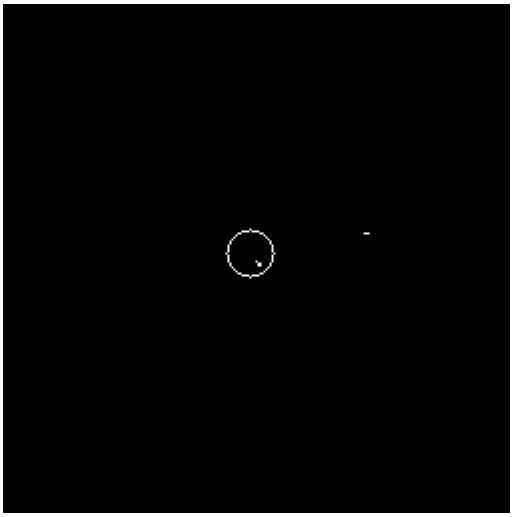
Figure 6.1 Database case *mdb209* - Malignant ROI with all microcalcifications in the cluster detected



(a) Original image

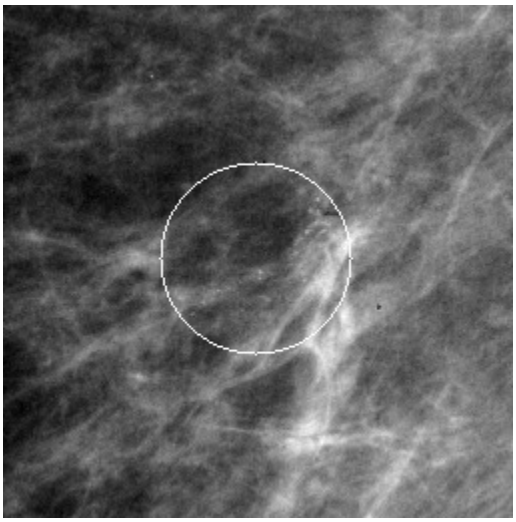


(b) Enhanced image after linearly combining the output of wavelet analysis, Median and Gaussian filtering

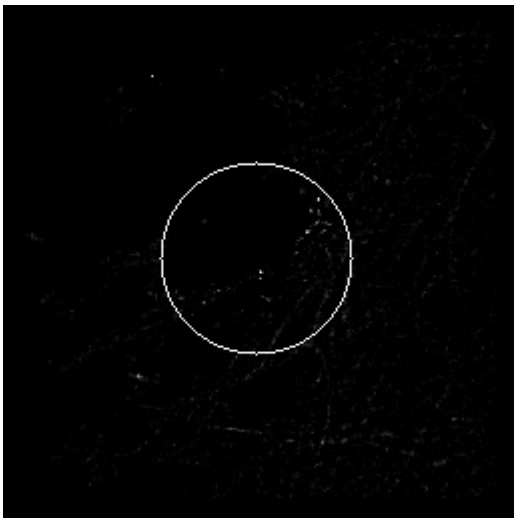


(c) represents the final image, after thresholding and applying all the post-processing operations

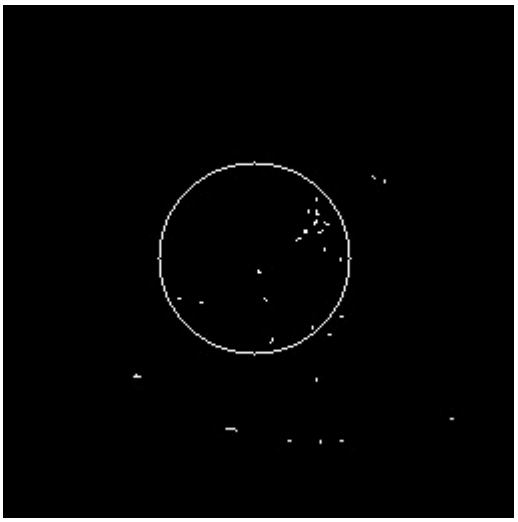
Figure 6.2 Database case *mdb218* - Benign ROI with microcalcification detected



(a) Original image



(b) Enhanced image after linearly combining the output of wavelet analysis, Median and Gaussian filtering



(c) represents the final image, after thresholding and applying all the post-processing operations

Figure 6.3 Database case *mdb231* - Malignant ROI with microcalcification cluster detected as well as some False Positives

Table 6.2 presents the performance of the proposed model with regards to Sensitivity and Specificity measurements. The first row assesses the performance of the proposed model in detecting all microcalcifications, Malignant or benign, in a database containing 99 images of all pathology (Malignant+Benign+Normal). The second row presents the results of the proposed model's ability to detect malignant clusters in a database containing 27 malignant and benign cases.

Table 6.2 Sensitivity and Specificity results for Malignant and Normal images

	Sensitivity	Specificity
Normal/Abnormal ROIs	100	11
Malignant/Benign ROIs	57	40

The sensitivity results in Table 6.2 show that the proposed model positively identifies all malignant and benign calcifications according to the ground truth. The proposed model also scores above average in the discrimination between Malignant and benign microcalcifications. One weakness of the mdoel is that it falsely flags certain normal breast structures as calcifications, which gives the low specificity rate. The specificity rate can be overlooked under the context that this model is intended to maximize the detection rate of microcalcifications where they exist. It should be pointed out that in most images where microcalcifications are present, they are uniquely detected by the proposed model without falsely flagging other parts of the same image as positive.

As supported by the sensitivity results of the proposed model (Table 6.2), all microcalcification clusters are detected and stand out from the background to a significant degree. The proposed model in some cases falsely flags non-calcification objects as microcalcifications, which is the cause of the low specificity rate. This point is best illustrated in Fig. 6.3, where the microcalcification cluster is clearly distinguishable from the background after the filtering stage, even though the final thresholding process introduces some artifacts. A closer inspection reveals that the false positives in the thresholded image follow the path of the curvilinear structures in the original and enhanced image. The different filters and kernel dimension parameters used in this work had their unique side-effects in emphasizing certain curvilinear structures; their combination through scalar multiplication was intended to reduce the over-emphasis of those structures by individual filters. While the filters used significantly reduced the effect of Curvilinear structures in Fig. 6.3 as well as the other images, they are nevertheless still pronounced in this image, which could be the cause of the false positives. The challenges encountered during the determination of an optimal threshold include mammogram image contrast, breast density and curvilinear structures. These are common challenges that are not unique to this project [121]. The results verify that intensity alone is insufficient as a criteria for the proper segmentation of microcalcification objects.

Table 6.3 shows a comparison with related works. The proposed model performs significantly

better than related works in terms of sensitivity rates with regards to Normal/Abnormal ROIs with the highest score of 100%. However, it does not compare well in specificity rates, scoring 11%.

Table 6.3 Comparison with related works using the MIAS database

Author	Sensitivity (TP rate)	Specificity
Oh et al. [121]	93.1	87.5
Jian et al. [58]	83	
Mohanlin et al. [122]	96.55	60
Our work	**100**	**11**

Looking at the sensitivity value in Table 6.3, it can be concluded that the integration of the wavelet filters and the Laplace operators definitely contributed to the high detection rate for microcalcifications where present. This maximized the probability of positive pixels being identified, even if at the cost of falsely flagging non-calcification objects as positive. In practice, false negatives are more highly penalized than false positives [123]; false negatives imply delayed treatment as the alert is not raised, which might lead to a fatal prognosis for the patient. This supports the merits for this model in the sense that, its high sensitivity performance implies that calcifications are highly unlikely to be missed - the prompting of suspicious regions can be useful in contexts where such information is needed.

6.3 Feature extraction

In the second phasse, machine learning based on extracted features was used to improve on the high false positive rate reported in the previous section. The proposed model was comparatively analyzed with related work in the literature. The results are shown in Tables 6.4-6.5, benchmarked using the metrics discussed in the section 6.1. For a more clear perspective of the overall performance of the models for comparative assessment, the scores for all the metrics are averaged and presented in the last column of the results. The parameter values for k are taken from the set $[1,3,5,7,9,11]$; odd numbers were picked to avoid tie scenarios during voting. Related work in the literature has rarely gone above 11, which guided its choice as the maximum number of neighbors [44]; the experiments conducted in this work also have not shown improvement of results that would warrant consideration of values higher than 11 for k. For referential convenience, the first approach, containing Haralick and Geometric features is referred to as Model 1 (Table 6.4); the two-dimensional set comprising the SVM and QDA scores is referred to as Model 2 (Table 6.5) and the proposed model containing features derived from the classifier scores as Model 3 (Table 6.6) or simply “the proposed model”.

Table 6.4 Performance benchmark using selected Haralick and geometric features. PPV is the Positive Predictive Value benchmark, also known as Precision. The highlighted row marks the best performing k value based on average score of all metrics

k	Accuracy	PPV	Sensitivity	Specificity	**Average**
1	0.7784	0.1429	0.1739	0.8596	**0.4887**
3	0.8505	0.2000	0.08696	0.9532	**0.5227**
5	0.8608	0.1667	0.04348	0.9708	**0.5105**
7	0.8711	0.2500	0.04348	0.9825	**0.5368**
9	0.8763	0.3333	0.04348	0.9883	**0.5604**
11	0.8814	0	0	1	**0.4704**
Average	**0.8531**	**0.1822**	**0.0652**	**0.9591**	**0.5149**

Table 6.5 Performance benchmark using SVM and QDA scores only. PPV is the Positive Predictive Value benchmark, also known as Precision. The highlighted row marks the best performing k value based on average score of all metrics

k	Accuracy	PPV	Sensitivity	Specificity	**Average**
1	0.7474	0.1389	0.2174	0.8187	**0.4806**
3	0.799	0.1364	0.1304	0.8889	**0.4887**
5	0.8918	1	0.0870	1	**0.7447**
7	0.8918	1	0.0870	1	**0.7447**
9	0.8814	0	0	1	**0.4704**
11	0.8814	0	0	1	**0.4704**
Average	**0.8488**	**0.3792**	**0.0870**	**0.9513**	**0.5666**

Table 6.6 Performance benchmark for the derived feature set comprising Statistics on SVM and QDA scores. PPV is the Positive Predictive Value benchmark, also known as Precision. The highlighted row marks the best performing k value based on average score of all metrics

k	Accuracy	PPV	Sensitivity	Specificity	**Average**
1	0.9278	0.68	0.7391	0.9532	**0.8250**
3	0.9433	0.7727	0.7391	0.9708	**0.8565**
5	0.9485	0.8095	0.7391	0.9766	**0.8684**
7	0.9536	0.8182	0.7826	0.9766	**0.8828**
9	0.9485	0.76	0.8261	0.9649	**0.8749**
11	0.9485	0.76	0.8261	0.9649	**0.8749**
Average	**0.9450**	**0.7667**	**0.7754**	**0.9678**	**0.8637**

According to the results in Table 6.4, Model 1 performs the least in sensitivity and PPV, but strongly in specificity (95.91%), with the best performance attained at $k = 9$. Notably at this point, the high specificity result is a pattern shown by all models with scores at 95.13% and 96.78% for Model 2 and Model 3 respectively. This model performs poorly at all parameter values of k in the PPV criteria, with an average precision of 18.22%.

Model 2 (Table 6.5) marginally improves the sensitivity score at 8.7% in comparison to Model 1. Its PPV score is more than double that of Model 1, but is still significantly lower at 37.92% when averaged across all values of k. However, it should be noted that it gives perfect scores for the PPV metric at $k = 5$ and $k = 7$. Evidently, these two parameter values offer the best performance for this model. It outscores Model 1 in all benchmarks with an average improvement of 18.43%, considering the optimal parameter settings for both models. Its best sensitivity score of 8.7% (at its optimal k value, or 21.74% across all values of k) is however below ideal for practical application.

The proposed model (Table 6.6) outperforms both Model 1 and Model 2 in the average scores of all metrics. It has the best performance at $k = 7$ with an average score of 88.28% across all metrics, which is a significant improvement by 13.81% and 32.24% over Model 2 and Model 1 respectively. It registers its least performance of 68% in the PPV metric at $k = 1$. Of note, especially in comparison with the other models, is its consistent performance across all benchmarks, with scores significantly above 50%.

The high specificity scores, which are perfect in some settings for Model 1 and Model 2, might initially suggest good discrimination capability regarding negative cases for those models, until they are balanced with the other metrics. The experimental setup involved using all database images as query images at some point in the iterations. Given that the dataset samples are skewed towards non-malignant cases, the high specificity values of Model 1 and Model 2 might have undesirably been buoyed by that fact. This reasoning is supported by their low sensitivity and precision values. The high sensitivity values of the proposed model imply that it is relatively robust and effective in its ability to discriminate malignant cases even when the dataset's class ratio is significantly imbalanced.

Studies have discovered a high rate of false positives as well as missed detections by radiologists during breast cancer screening, with estimates placing the radiologists' sensitivity at about 75% [124]. Accurate query results based on visual content have been reported to be a significant diagnostic aid to radiologists [36, 84]. The significantly positive scores of the proposed model across all the metrics employed imply a high and consistent ability to extract database cases closely matching a given query sample; the incorporation of more suitable features as demonstrated by the proposed model can enhance the accuracy of CBIR-based CAD systems in breast cancer diagno-

sis. Such systems when used as a second opinion, can in turn improve the quality of diagnostic decisions [36,110]. The experimental results show that the proposed model can retrieve the correct results for 78 out of 100 queries involving positive samples. It retrieves the correct image results for 98 out of 100 queries involving negative samples. The high predictive value of 82% assigns a commensurate credibility to the model's positive results. Therefore, there is a high chance that the positive cases returned by the model are indeed positive. This is a desired attribute of a model in a clinical setting given that a low PPV leads to additional costs and negative pyschological effect on patients as follow up examination is necessitated to establish the actual diagnosis.

Most classification errors of the proposed model can be attributed to mammogram cases with dense fibroglandular tissues (Fig. 6.5); the intensity profile of such cases is very similar to that of microcalcifications making differentiation more challenging - this has been noted as a problem in similar studies as well and remains an open area of research [110, 125]. Our model nonetheless contributes to the field of computer aided diagnosis of breast cancer by introducing an improved feature characterization approach. Vijayalakshmi et al. [124] present a similar system combining the Local Binary Pattern (LBP) and the Artificial Neural Network classifier. While they report high scores (between $92.5\% - 100\%$) in the four metrics they use (specificity, accuracy, sensitivity and accuracy), their tests are however conducted on a smaller dataset of 80 images. Furthermore, the classes represented in their dataset are equally balanced as contrasted to the imbalanced class representation in this study.

Table 6.7 demonstrates the competitive performance of the proposed model in comparison to other related works. In particular, Tsochatzidis et al. [51] present a supervised retrieval model based on the SVM for malignancy assessment. Their feature vector is derived from participation values of support vectors. Their model considers all BI-RADS categories with the database composed of a total of 87 ROIs. In Table 6.7, we also consider their average accuracy score across all categories. Fig. 6.4 and 6.5 give image results to sample queries for both accurate and inaccurate scenarios respectively. In Fig. 6.4, the inaccurate result appears in fifth position among the returned results. In Fig. 6.5, the inaccurate results appear in positions one, two, five and six.

Table 6.7 Comparison of proposed model against recent similar works. The values are expressed as percentage scores

Author	PPV(%)	Specificity(%)	Sensitivity(%)	Accuracy(%)
Tsochatzidis et al. [51]	N/A	N/A	N/A	60
Our model	**82**	**98**	**78**	**95**

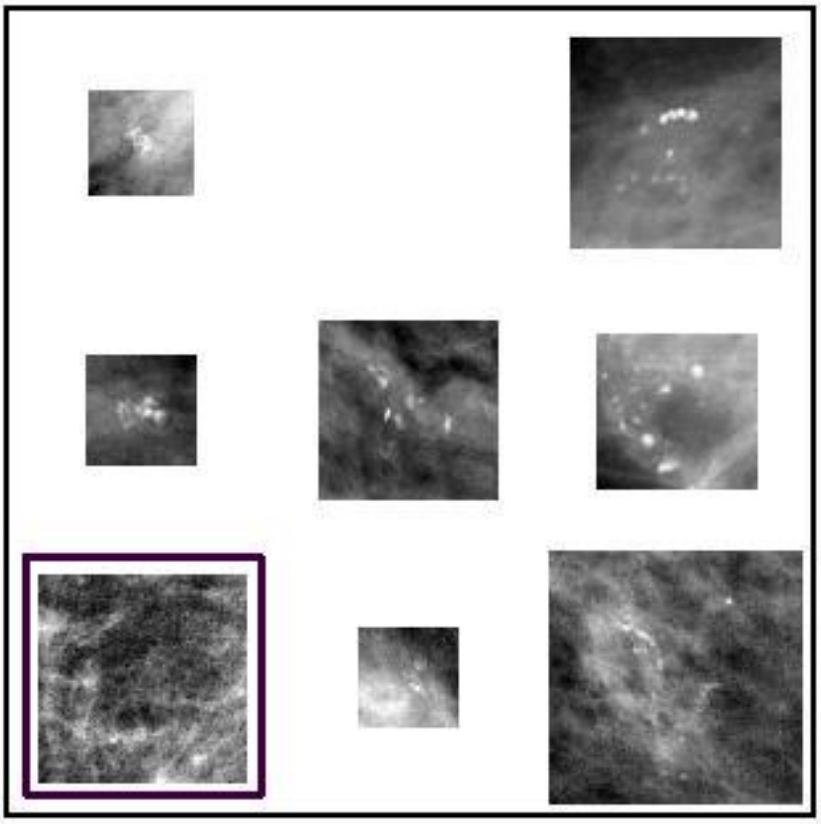


Figure 6.4 Accurate query results using image mdb227 ROI. The query image appears on the top-left. The incorrect result has been highlighted by a dark rectangle (bottom left).

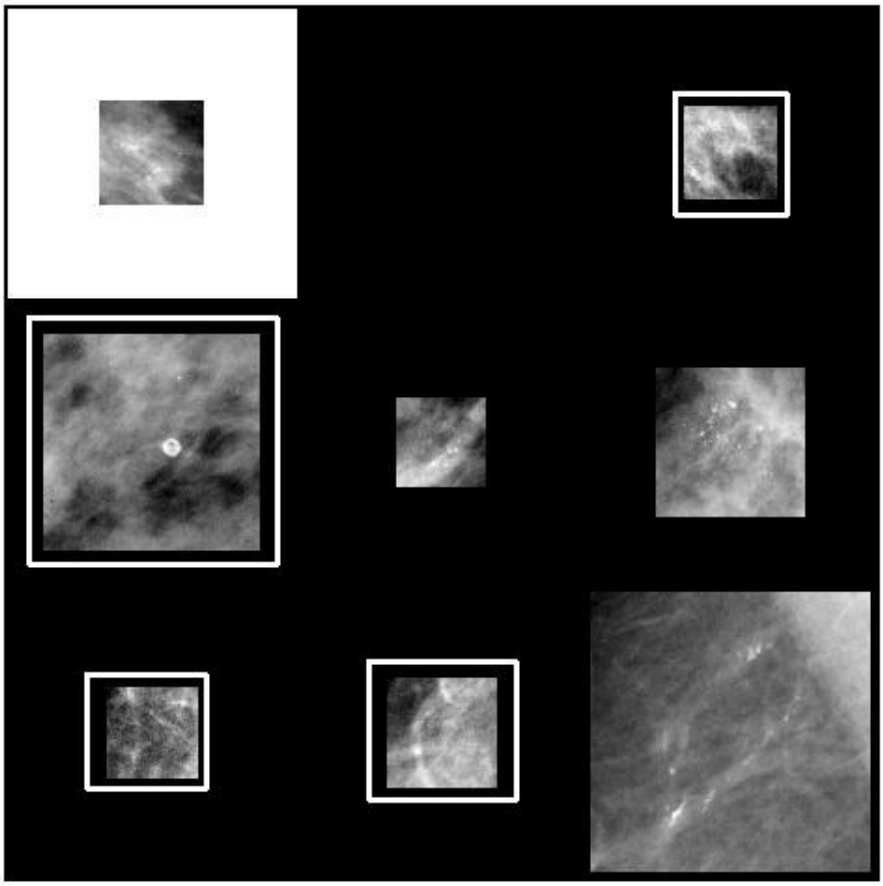


Figure 6.5 Inaccurate query results using an mdb238 ROI. The query image appears on the top-left. The incorrect results have been highlighted by a white rectangle

6.4 Parallel extraction of features

The development tools used in this study included gcc-5.3.1 (C++11), OpenCV 3.0 and OpenMPI 1.8.8. The system was tested on a cluster containing up to 20 homogenous nodes, each having 24 Intel 5th generation CPUs and 128*GB* of memory. It should be noted that a thread in the context of this paper implies a physical processor because a mapping of one thread to one processor was enforced; the terms can therefore be used interchangeably and should not cause ambiguity to the reader. Due to resource limitation (maximum 240 nodes available to this study due to fair use policy), there were two scenarios for the hardware setup specific to a given experiment: the first was a logical setup that divided available nodes to give 60 nodes, each having 4 processors. The second scenario considered 20 nodes having 12 processors each. The hardware was set up to ensure that a physical compute node was considered only once, in order to force inter-node cooperation, i.e., ensure every MPI communication call actually involved at least two disparate hardware nodes. The performance of the algorithm was benchmarked on the running time T_p, Speedup S_p and Efficiency E, which are calculated as follows,

$$S_p = \frac{T_1}{T_p} \tag{6.5}$$

$$E = \frac{S_p}{p} = \frac{T_1}{pT_p} \tag{6.6}$$

T_1 is the time taken to complete task execution in the serial version of the program, as measured by the wallclock time. p the number of processes and T_p the overall time taken from the start of computation to the time the last process terminates. Ideally, it would be desired to have a speedup (S_p) equal to p which, however, is rare in practice (implying $S_p \leq p$) due to the overheads mentioned previously. We can summarize all the overheads (including $T(n)$) incurred in our parallel system as the total overhead or overhead function T_o, defined as $T_o = pT_p - T_s$, where T_s is the amount of time spent on the actual computation. Efficiency measures how far from an ideal system the chosen model performs at, and ranges between 0 and 1. An ideal system would give an efficiency of $E = 1$.

The models were tested on an increasing database size, [100,200,322] images to measure their scalability performance. Scalability measures how well a given model handles an increasing input size. First, the serial program was benchmarked using running time, with its performance measured against the increasing database size. The results were as follows,

- Database size=100, Running time=408.34s
- Database size=200, Running time=537.43s
- Database size=322, Running time=876.75s

The first experiment tested the scalability of the system with an increase in the number of compute nodes as well as the database size; the results are shown in Table 6.8. The node pool for this experiment ranged up to 60, with four cores at each node. Generally, the proposed model significantly improves the computational complexity of the serial model for every pair of variable considered (i.e. database size and node size). Even the minimum speedup achieved is significant when considered in isolation. For any given fixed database size, the model shows a general improvement with an increase in the number of compute nodes.

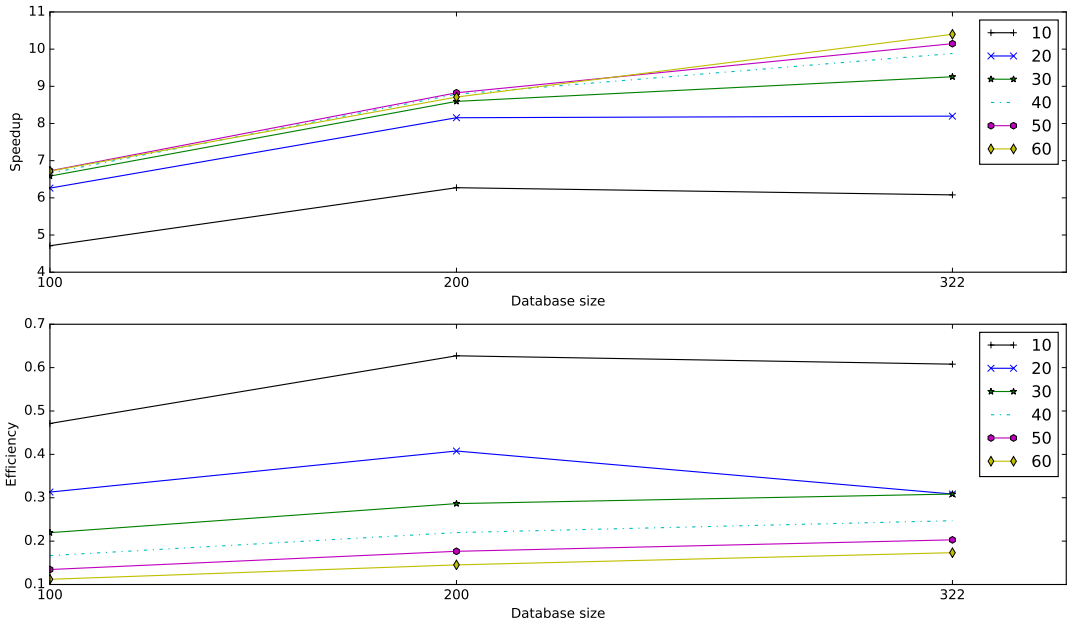


Figure 6.6 A plot of the proposed model's speedup/efficiency performance considering an increasing database size, at various node sizes. Generally, the model shows an increasing speedup and efficiency performance as the database size is increased.

Fig. 6.6 shows the graphs of the proposed model's performance considering speedup/efficiency vs database size for the node sizes $10, 20, 30, 40, 50$ and 60. Overally, there is an increase in the speedup performance of the model with an increasing database size. There is a bigger speedup

Table 6.8 Performance of the proposed model considering an increasing database size and node size. The number of threads was fixed at a constant value of 1.

DB Size	Running time (Seconds)	Node size	Speedup	Efficiency
100	86.662673	10	4.7118325095	0.4711832509
100	65.217121	20	6.2612392841	0.3130619642
100	62.013803	30	6.5846630951	0.2194887698
100	61.273297	40	6.664240705	0.1666060176
100	60.709272	50	6.726155438	0.1345231088
100	60.777712	60	6.7185813115	0.1119763552
200	85.678197	10	6.2726576751	0.6272657675
200	65.903529	20	8.1547985086	0.4077399254
200	62.533336	30	8.5942960088	0.2864765336
200	61.158355	40	8.7875156224	0.2196878906
200	60.894144	50	8.8256433985	0.176512868
200	61.697075	60	8.710785722	0.145179762
322	144.212315	10	6.0795778779	0.6079577878
322	106.950759	20	8.1976977835	0.4098848892
322	94.720359	30	9.2561938031	0.3085397934
322	88.714526	40	9.8828234736	0.2470705868
322	86.407233	50	10.146720009	0.2029344002
322	84.307508	60	10.3994296688	0.1733238278

degree considering a bigger number of nodes (60 in this case) compared to a smaller node size (such as 10). A significant trend shown in the results is the relatively higher speedup values as the database size increases. This implies a higher probability of a performance gain as the input space increases, which is an underlying fact of medical databases; Medical databases grow in size due to continued capturing of patient data as well advances in technology, which demand more storage space as well as imposing an additional processing burden [19]. A big database is highly beneficial to CBMIR systems since it provides more information for radiologists to base their diagnostic decision, thereby enhancing the quality of the same.

The efficiency of the model reports a decrease with an increase in the number of nodes for a fixed database size (Table 6.8). This can probably be explained by a higher cumulative idling time implied by a bigger node pool as compared to the database size. This is evident in the trend that an increase in the database size gives a proportionate increase in the efficiency of the model (Fig. 6.6). For instance, the efficiency increases by 30% when the database size is increased by 222

images. The efficiency for a smaller node size ($N = 10$) is relatively higher than for the largest node size ($N = 60$) because of a lesser communication penalty involved in coordinating fewer nodes. The algorithm shows an increasing efficiency when the database size is scaled upwards, which implies that it leverages well the existing resources to handle an increasing input size. The coarse partitioning at the inter-node level for a smaller database size means that the tasks are skewed to some nodes at this stage, but this is corrected as the database increases in size. In practice, the ability of an algorithm to be more efficient or retain efficiency with an increasing database size is desired, given that medical databases increase in size as more patient data is captured both for monitoring, research and educational purposes. Since medical databases are voluminous in most practical cases [16, 17], it is therefore positively significant that the algorithm shows a relatively higher efficiency at higher database sizes.

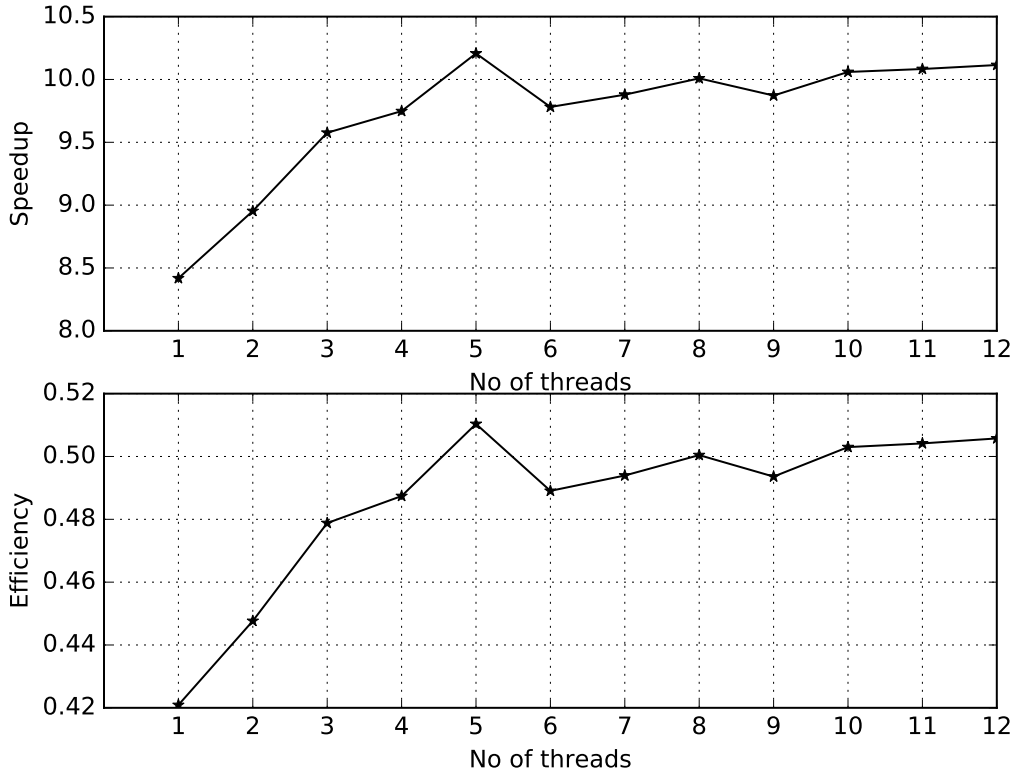


Figure 6.7 Performance of the proposed model with monotonic increase of threads/processor cores. The Database and Node size are kept constant at 300 and 20 respectively. Speedup values indicate the factor by which the parallel model is faster than the serial version, while efficiency is a value between 0 (worst) and 1 (best).

An experiment was also conducted to test the effect (and its magnitude) to the proposed model's performance as more processor cores (or threads) are added to the compute pool. The node pool

for this experiment was reduced to give a wider range of threads in order to make their effect more pronounced. Fig. 6.7 shows plots of the speedup and efficiency values. The speedup degree increases with an increase in the number of processor cores with a local maximum at 5 cores, after which it tapers off with minor fluctuations. Similarly to the node size, it is desirable that a parallel model scales well with an increasing size of cores as demonstrated by the proposed model. The efficiency results show a similar trend to the speedup values, depicting better use of resources with an increase in the number of cores. Based on these results, the model can therefore be described to scale well as more cores are added to the nodes. Since the task is divided coarsely along image boundaries, a higher k value would be appropriate for a proportionally higher number of cores, with an image being mapped to a particular core. A block of images can therefore be assigned to a particular node to offset latency costs, then assigned to individual cores once at the node. The relatively lower latency penalty among the symmetric multiprocessors (SMPs) at the node level should imply a lower overall latency cost for the entire system. This does not preclude the negative impact of other probable penalties such as thread management costs. There should be a positive impact on performance where the inter-node latency cost significantly outweighs the other costs.

The final experiment compared the effect of manually assigning the $k-$ value vis-a-vis automatically setting it according to performance analysis (Eq. 5.19). This experiment had a similar setup to the second experiment, with 20 nodes having 12 processors each. The manual k values were taken from the integer range $k \in [1, 12]$. The various manually selected k values give various speedup and efficiency performances. The automatically calculated method sustains the k value at a near optimal level, which ensures a stable and nearly optimal performance. This experiment demonstrated the speedup potential of factoring in the communication latency cost during task assignment to improve the computational complexity of a parallel model.

As further work, the proposed model can possibly be enhanced for improved accuracy by including a physician-in-the-loop approach, where the results are assigned relevance scores, with the same being used to modify the weights of the attributes. Additionally, the incorporation of other pathological features important to the detection of breast cancer such as breast masses can be included to provide a holistic diagnostic approach, given that this study focused only on microcalcifications. Modeling of the dense tissues and their differentiation from microcalcifications can be investigated as a means of reducing the false positive cases encountered in this study. Regarding run-time performance, while the proposed model shows increasing efficiency with an increasing database size, further work can address making it more efficient in the case of a higher node size as compared to the database size.

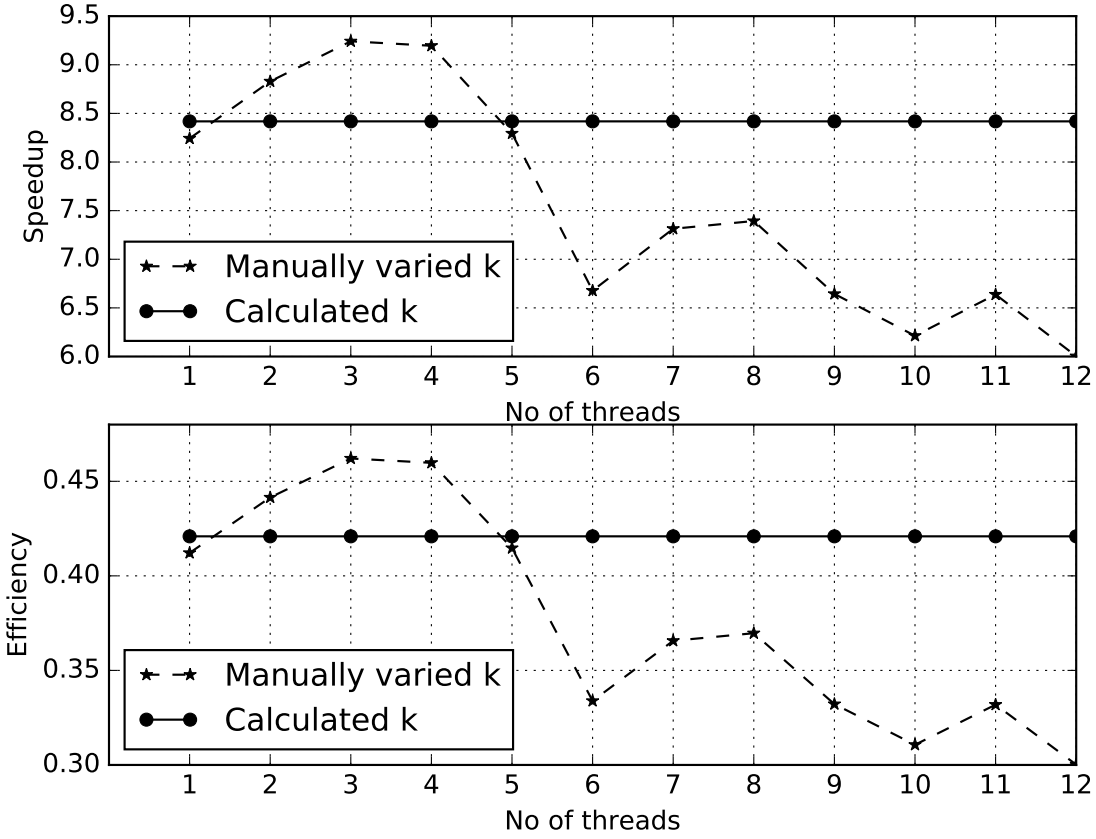


Figure 6.8 Comparison of the proposed model's performance when k is manually varied incrementally and when it is calculated according to Eq. 5.19. The database size is fixed at 322, number of threads=1

Chapter 7

Conclusion and future work

In this thesis, we presented an approach for improving the accuracy and responsiveness of mammogram retrieval by optimizing various components of the system. The retrieval was based on similarity with regards to pathology and specifically, microcalcifications. We first looked at the optimal detection of microcalcifications in digital mammogram images. We also investigated the discrimination between Malignant and benign subclasses of microcalcifications. To this end, the wavelet and Laplace filters were optimally integrated to amplify microcalcifications followed by postprocessing to reduce the number of false positives. Each of the filters have their unique strengths and side-effects in detection of calcification-like objects; the essence was to combine them optimally to highlight their strengths while canceling their side-effects. The combined filter model detected all present calcifications, with a sensitivity rate of 100%, in all mammograms as demarcated by expert radiologists based on accompanying ground truth. The false-positive rate was significantly higher based on the lower specificity rate, a factor that can be investigated through use of efficient feature extraction and classification methods to characterize calcifications and curvilinear structures to reduce the false-positive rate. In accordance with the objectives of the study, the proposed model demonstrated the effectiveness of combining the likelihood maps from different filters in improving detection of calcification objects.

The second part of this thesis presented an improved model for the retrieval of mammograms based on their pathology. The main focus and contribution was the combination of classifier scores, with statistics on the same as a means of improving the accuracy of the retrieval of mammogram cases based on pathology and more specifically, the malignancy of clusters, where present. The feature characterization model was further improved by appropriate weighting of the features based

on results of their individual discrimination ability. Experiments benchmarked the model's performance using a wide range of metrics, with results showing increased relative effectiveness of the proposed model over the common application of texture/geometric features or their scores alone as widely applied in the literature. The scores achieved were 95%, 82%, 78%, and 98% on the accuracy, positive predictive value, sensitivity and specificity benchmarks respectively. Further works on this model can consider extending the classification problem to include other pathologies according to BI-RADS classes and addressing the negative effect of dense fibroglandular tissues on microcalcification characterization.

The final phase leveraged cluster and multi-core parallel computing to the task of feature extraction with the aim of reducing the computational cost of the process. Both models were based on Single Instruction Multiple Data (SIMD) model with the tasks assigned dynamically in a Master-Slave approach. In the model, the master node was dedicated to task assignment and freed from computations in order to reduce idling among worker/compute nodes. Worker nodes were served with whole images and execute all the subtasks individually. A method was proposed for optimizing the number of tasks that are assigned to any given node, considering the communication cost and other penalties. The models were benchmarked by the speedup degree and efficiency metrics. According to the experimental results, the model achieved speedup values of between $4.7x$ and $10.4x$, and efficiency values of between 0.11 and 0.62. The number of nodes as well as the database size were varied to measure the scalability of the model. The results expectedly showed an improvement in all benchmarks by all parallel models over the serial model. Specifically, addition of more nodes improved the execution time with a significant speedup. Efficiency was found to deteriorate with further increase of nodes for a fixed database, but did not degrade with an increase in the number of cores.

The model proposed in this work can be used as a basis in the development of a CBIR-based CAD system that would help improve the radiologists' accuracy in the diagnosis of a case in hand, by making reference to similar historical cases. There exist a number of areas for improvement on the proposed model in further works; in this project, classifier training using ground-truth was used to model the radiologist's perception of similarity, as well as bridging the semantic gap in the description of microcalcifications. While that proved effective, there is a potential of further refining the performance of the proposed model through use of relevance feedback rounds as has been tried in related works. Due to resource constraints, this study did not consider the GPU, which has potential of further speeding up the computations of especially data parallel processing operations. The partitioning problem considered only multiples of whole images, future work will consider further subdivision of a given image for concurrent processing where possible.

Bibliography

[1] T. Jabid, H. Kabir, and O. Chae, "Gender Classification using Local Directional Pattern (LDP)," In *International Conference on Pattern Recognition*, pp. 2162–2165 (2010).

[2] F. Valente, C. Costa, and A. Silva, in *Content Based Retrieval Systems in a Clinical Context*, O. F. Erondu, ed., (2013), Chap. 1.

[3] L. B. Holder, I. Russell, Z. Markov, A. G. Pipe, and B. Carse, "Current And Future Trends In Feature Selection And Extraction For Classification Problems," International Journal of Pattern Recognition and Artificial Intelligence **19,** 133–142 (2005).

[4] M. C. Oliveira, W. Cirne, and P. M. de Azevedo Marques, "Towards Applying Content-based Image Retrieval in the Clinical Routine," Future Generation Computer Systems **23,** 466–474 (2007).

[5] M. Brady, D. Gavaghan, R. Highnam, A. Knox, S. Lloyd, A. Simpson, and D. Watson, "Grid Computing For Digital Mammography," In *UK e-Science All Hands Meeting*, (eDiaMoND, 2003).

[6] N. Perez, M. A. Guevara, M. Vaz, R. Ramos, M. Rubio, and F. Castrillo, "A CAD tool for mammography image analysis. Evaluation on GRID environment," IJCSI International Journal of Computer Science Issues **10,** 255–259 (2007).

[7] M. C. Oliveira, W. Cirne, and P. M. A. Marques, "Towards applying content based image retrieval in the clinical routine," IEEE Trans. Image Processing **11,** 467–476 (2002).

[8] M. H., N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems," In *Medical applications-clinical benefits and future directions*, pp. 1–23 (International Journal of Medical Informatics, 2004).

[9] N. Strickland, "PACS (picture archiving and communication systems): filmless radiology," Archives of Disease in Childhood **83,** 82–86 (2000).

[10] P. Welter, J. Riesmeier, B. Fischer, C. Grouls, C. Kuhl, and T. M. Deserno, "Bridging the integration gap between imaging and information systems: a uniform data concept for content-based image retrieval in computer-aided diagnosis," Journal of the American Medical Informatics Association **18,** 506–510 (2011).

[11] Rosenthal David F, Bos JoAnne M, Sokolowski Rachael A, Mayo Jennifer B, Quigley Kerry A, Powell Roger A, and Teel Mary-Marshall, "A Voice-enabled, Structured Medical Reporting System," Journal of the American Medical Informatics Association **4,** 436–441 (1997).

[12] T. Freer and M. Ulissey, "Challenges of medical image processing," Radiology **220,** 781–786 (2004).

[13] J. C. Felipe, M. X. Ribeiro, E. P. M. Sousa, A. J. M. Traina, and C. Trainan, "Effective Shape-based Retrieval and Classification of Mammograms," In *SAC '06 Proceedings of the 2006 ACM symposium on Applied computing*, pp. 250–255 (Association for Computing Machinery, 2006).

[14] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for Automated Detection and Classification of Masses in Mammograms," Pattern Recognition **39,** 646–668 (2006).

[15] G. G. *et al.*, "Downgrading BIRADS 3 to BIRADS 2 category using a computer-aided microcalcification analysis and risk assessment system for early breast cancer," Computers in Biology and Medicine **40,** 853–859 (2010).

[16] I. Scholl, T. Aach, T. M. Deserno, and T. Kuhlen, "Challenges of medical image processing," Journal of Computer Research and Development **26,** 5–13 (2011).

[17] P. M. de Azevedo-Marques and R. M. Rangayyan, in *Content-based Retrieval of Medical Images: Landmarking, Indexing and Relevance Feedback*, *Synthesis lectures on Biomedical Engineering*, 48 ed., J. D. Enderle, ed., (Morgan and Claypool, 2013).

[18] D. Storcheus, A. Rostamizadeh, and S. Kumar, in *Journal of Machine Learning Research: Workshop and Conference Proceedings*, N. D. Lawrence, ed., (2015), Vol. 44, pp. 1–18.

[19] Luo Jake, Wu Min, Gopukumar Deepika, and Zhao Yiqing, "Big Data Application in Biomedical Research and Health Care: A Literature Review," Biomedical Informatics Insights **8,** 1–10 (2015).

[20] V. H. Kondekar, V. S. Kolkure, and S. Kore, "Retrieval Techniques based on Image Features: A State of Art approach for CBIR," International Journal of Computer Science and Information Security **7,** 6753–6758 (2010).

[21] T. Deselaers, D. Keysers, and H. Ney, "Features for Image Retrieval: A Quantitative Comparison," In *26th DAGM Symposium on Pattern Recognition (DAGM 2004)*, **3175,** 228–236 (Lecture Notes in Computer Science, 2004).

[22] D. A. Manolescu, "Feature Extraction-A Pattern for Information Retrieval," In *Proceedings of the 5th Pattern Languages of Programming*, pp. 1–18 (1999).

[23] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Transactions On Pattern Analysis And Machine Intelligence 22 (2000).

[24] S. R. Sternberg, "Biomedical Image Processing," Computer **16,** 22–34 (1983).

[25] M. O. Guld, C. Thies, B. Fischer, and T. M. Lehmann, "A generic concept for the implementation of medical image retrieval systems," International journal of medical informatics **76,** 252–259 (2007).

[26] E. S. Burnside, "Use of Microcalcification Descriptors in BI-RADS 4th Edition to Stratify Risk of Malignancy," Radiology **242,** 388–395 (2007).

[27] J. Ren, "ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging," Knowledge-Based Systems **26,** 144–153 (2012).

[28] N. C. Institute, "Breast cancer,", online (2011).

[29] B. A. Kohler, E. Ward, B. J. McCarthy, M. J. Schymura, L. A. G. Ries, C. Eheman, A. Jemal, R. A. Anderson, U. A. Ajani, and B. K. Edwards, "Report to the Nation on the Status of Cancer," Journal of the National Cancer Institute pp. 1975–2007 (2015).

[30] A. Wróblewska, P. Boniński, A. Przelaskowski, and M. Kazubek, "Segmentation and feature extraction for reliable classification of microcalcifications in digital mammograms," Opto - Electronics Review **11,** 227–235 (2003).

[31] N. C. Institute, "what you need to know about breast cancer,", Online (2009).

[32] R. Smithuis and R. Pijnappel, "Breast - Calcifications Differential Diagnosis,", Online, 2008.

[33] C.-T. Li, C.-H. Wei, C.-H. Wei, and C.-T. Li, "A Content-Based Approach to Medical Image Database Retrieval," In *Database Modeling for Industrial Data Management: Emerging Technologies and Applications*, (Idea Group Publishing, 2005).

[34] J. E. E. de Oliveira, A. M. C. Machado, G. C. Chávez, A. P. B. Lopes, T. M. Deserno, and A. d. A. Araújo, "MammoSys: A content-based image retrieval system using breast density patterns," Computer Methods and Programs in BioMedicine **99,** 289–297 (2010).

[35] S. Dreiseitl and M. Binder, "Do physicians value decision support? A look at the effect of decision support systems on physician opinion," Artificial Intelligence in Medicine **33,** 25–30 (2005).

[36] J. F. Gilbert, S. M. Astley, M. G. Gillan, O. F. Agbaje, M. G. Wallis, J. James, C. R. M. Boggis, and S. W. Duffy, "Single Reading with Computer-Aided Detection for Screening Mammography," The New England Journal of Medicine **359,** 1675–1684 (2008).

[37] R. Nakayama, R. Watanabe, K. Namba, K. Takeda, K. Yamamoto, S. Katsuragawa, and K. Doi, "Computer-aided diagnosis scheme for identifying histological classification of clustered microcalcifications by use of follow-up magnification mammograms," Academic radiology **13,** 1219–1228 (2006).

[38] P. Singh, S. Singh, and G. Kaur, "A Study of Gaps in CBMIR using Different Methods and Prospective," World Academy of Science, Engineering and Technology **46,** 492–496 (2008).

[39] J. E. E. de Oliveira, A. d. A. Araújo, and T. M. Deserno, "Content-based image retrieval applied to BI-RADS tissue classification in screening mammography," World Journal of Radiology **3,** 24–31 (2011).

[40] A. Baldi, R. Murace, E. Dragonetti, M. Manganaro, O. Guerra, S. Bizzi, and L. Galli, "Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions," BioMedical Engineering OnLine 8 (2009).

[41] O. Trelles, "On the parallelisation of bioinformatics applications," Briefings in Bioinformatics **2,** 181–194 (2001).

[42] B. Zheng, "Computer-Aided Diagnosis in Mammography Using Content-based Image Retrieval Approaches: Current Status and Future Perspectives," Algorithms **2,** 828–849 (2009).

[43] S. K. Kinoshita, P. M. de Azevedo-Marques, R. R. Pereira, J. A. H. Rodrigues, and R. M. Rangayyan, "Content-based Retrieval of Mammograms Using Visual Features Related to Breast Density Patterns," Journal of Digital Imaging **20,** 172–190 (2007).

[44] M. E. Osman, M. A. Wahed, A. S. Mohamed, and Y. M. Kadah, "Computer Aided Diagnosis System for Classification of Microcalcifications in Digital Mammograms," In *26th National Radio Science Conference*, (2009).

[45] K. Arai, I. N. Abdullah, and H. Okumura, "Automated Detection Method for Clustered Microcalcification in Mammogram Image Based on Statistical Textural Features," International Journal of Advanced Research in Artificial Intelligence **1,** 22–26 (2012).

[46] A. Tieudeu, C. Daul, A. Kentshop, P. Graebling, and D. Wolf, "Texture-based analysis of clustered microcalcifications detected on mammograms," Digital Signal Processing **22,** 124–132 (2011).

[47] S. Wiesmuller and D. A. Chandy, "Content based mammogram retrieval using Gray Level Aura Matrix," In *International Joint Journal Conference on Engineering and Technology(IJJCET)*, pp. 217–221 (2010).

[48] C.-H. Wei, C.-T. Li, and R. Wilson, "A General Framework for Content-Based Medical Image Retrieval with its Application to Mammograms," In *SPIE*, **5748,** 134–143 (Medical Imaging 2005: PACS and Imaging Informatics, 2005).

[49] C.-H. Wei, Y. Li, and C.-T. Li, "Effective Extraction of Gabor Features for Adaptive Mammogram Retrieval.," In *International Conference on Multimedia and Expo*, pp. 1503–1506 (IEEE, 2007).

[50] V. D. Dhale, A. R. Mahajan, and U. Thaku, "A Survey of Feature Extraction Methods for Image Retrieval," International Journal of Advanced Research in Computer Science and Software Engineering **2,** 1–8 (2012).

[51] L. Tsochatzidis, K. Zagoris, M. Savelonas, N. Papamarkos, I. Pratikakis, N. Arikidis, and L. Costaridou, "Microcalcification oriented content-based mammogram retrieval for breast cancer diagnosis," In *IEEE International Conference Imaging Systems and Techniques (IST)*, pp. 257–262 (2014).

[52] L. Wei, Y. Yang, and R. M. Nishikawa, "Microcalcification classification assisted by content-based image Retrieval for breast cancer diagnosis," Pattern Recognition **42,** 1126–1132 (2009).

[53] L. Wei, Y. Yang, R. M. Nishikawa, and M. N. Wernick, "Learning of Perceptual Similarity from Expert Readers for Mammogram Retrieval," In *Biomedical Imaging: Nano to Macro*, pp. 1356–1359 (2006).

[54] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography," IEEE Transactions On Medical Imaging 23 (2004).

[55] M. Rizzi, M. D'Aloia, and B. Castagnolo, "Computer Aided Detection Of Microcalcifications In Digital Mammograms Adopting A Wavelet Decomposition," Integrated Computer-Aided Engineering **16,** 91–103 (2009).

[56] M. Rizzi and M. D. a. B. Castagnolo, "A Fully Automatic System for Detection of Breast Microcalcification Clusters," Medical and Biological Engineering **30,** 181–188 (2009).

[57] M. Rizzi and M. D. a. B. Castagnolo, "Detection of Microcalcifications in Digital Mammograms using the Dual-tree Complex Wavelet Transform," International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications **17,** 49–63 (2009).

[58] W. Jian, X. Sun, and S. Luo, "Computer-Aided Diagnosis Of Breast Microcalcifications Based On Dual-Tree Complex Wavelet Transform," BioMedical Engineering OnLine (2012).

[59] V. Alarcon-Aquino, O. Starostenko, J. M. Ramirez-Cortes, R. Rosas-Romero, J. Rodriguez-Asomoza, O. J. Paz-Luna, and K. Vazquez-Muioz, "Detection of Microcalcifications in Digital Mammograms using the Dual-tree Complex Wavelet Transform," International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications **17,** 49–63 (2009).

[60] T.Balakumaran, I.Vennila, and C. Shankar, "Detection Of Microcalcification In Digital Mammograms Using One Dimensional Wavelet Transform," International Journal of Computer Science and Information Security pp. 99–104 (2010).

[61] T.Balakumaran, I.Vennila, and C. Shankar, "Detection of Microcalcification in Mammograms Using Wavelet Transform and Fuzzy Shell Clustering," International Journal of Computer Science and Information Security **7,** 121–125 (2010).

[62] Y. G. Garud and N. G. Shahare, "Detection of microcalcifications in digital mammogram using wavelet analysis," American Journal of Engineering Research **02,** 80–85 (2013).

[63] N. B. Hamad, K. Taouil, and M. S. Bouhlel, "Mammographic Microcalcifications Detection using Discrete Wavelet Transform," International Journal of Computer Applications **64,** 17–22 (2013).

[64] N. B. Hamad, A. Masmoudi, and K. Taouil, "Wavelets Study for better multiresolution analysis in CAD of Microcalcification," International Journal of Computer Science **10,** 372–378 (2013).

[65] G. Jinghuan, C. Shenglai, G. Ku, and S. Zhaoqian, "Mammogram Image Classification Using Wavelet Based Haralick Features," International Journal of Signal Processing, Image Processing and Pattern Recognition **7,** 203–212 (2014).

[66] M. N. Arani and H. Ghassemian, "A Hierarchical Content-Based Image Retrieval Approach to Assisting Decision Support in Clinical Dermatology," Iranian journal of electrical and computer engineering **9,** 23–33 (2010).

[67] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowledge Data Engineering **17,** 491–502 (2005).

[68] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-Based Image Retrieval in Radiology: Current Status and Future Directions," Digital Imaging **24,** 208–230 (2011).

[69] I. El-Naqa, Y. Yang, N. P. Galatsanos, and W. N. Wernick, "Content-Based Image Retrieval For Digital Mammography," In *ICIP*, **3,** 141–144 (2002).

[70] Y. Y. H. Jing and R. M. Nishikawa, "Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer," Medical Physics **39,** 676–685 (2012).

[71] H. Jing, Y. Yang, and R. M. Nishikawa, "Regularization in Retrieval-driven Classification of Clustered Microcalcifications for Breast Cancer," Journal of Biomedical Imaging **2012,** 3:1–3:8 (2012).

[72] M. Rahman, P. Bhattacharya, and B. C. Desai, "A Framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," IEEE Trans On Information Tech. in Biomedicine 11 (2007).

[73] L. Tsochatzidis, K. Zagoris, M. Savelonas, and I. Pratikakis, "SVM-based CBIR of Breast Masses on Mammograms," In *Proceedings of the 3rd International Conference on Artificial Intelligence and Assistive Medicine - Volume 1213*, 14 **26,** 26–30 (CEUR-WS.org, Aachen, Germany, Germany, 2014).

[74] C. Wen-hao, F. Yu-Chun, Y. Ji-feng, and Z. wu, "Multi-core based parallel computing technique for content-based image retrieval," Journal Of Shanghai University (English Edition) **14,** 55–59 (2010).

[75] B. Zheng, Y. H. Chang, X. H. Wang, W. F. Good, and D. Gur, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," Academic Radiology **6,** 327–332 (1999).

[76] C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. Aisen, and L. Broderick, "Local versus Global Features for Content-Based Image Retrieval," IEEE Workshop on Content-Based Access of Image and Video Libraries (1998).

[77] P. S. Rao, E. V. G. Samuel, V. Prasad, R. Ravikanth, and S. Varma, "An Approach for CBIR System Through Multi Layer Neural Network,", 2010.

[78] M. Emmanuel, D. R. Babu, J. Jagdale, P. Game, and G. P. Potdar, "Parallel Approach for Content Based Medical Image Retrieval System," Journal of Computer Science **6,** 1258–1262 (2010).

[79] H. Smita, G. Monika, and C. Shraddha, "Content Based Image Retrieval Using HADOOP Map Reduce," International Journal of Computer Science Trends and Technology **2,** 1088–1089 (2014).

[80] S. Jai-Andaloussi, A. Elabdouli, A. Chaffai, N. Madrane, and A. Sekkaki, "Medical content based image retrieval by using the Hadoop framework," In *International Conference on Telecommunications*, **20,** 1–5 (2013).

[81] K. Yadav, A. Srivastava, A. Mittal, and M. Ansari, "Parallel Implementation of Shape based Image Retrieval Approach on CUDA in Compressed Domain," International Journal of Computer Applications Special Issue on Novel Aspects of Digital Imaging Applications pp. 15–22 (2011).

[82] H. Heidari, A. Chalechale, and A. A. Mohammadabadi, "Parallel Implementation of Color Based Image Retrieval Using CUDA on the GPU ," I.J. Information Technology and Computer Science pp. 33–40 (2014).

[83] S. F. da Silva, M. X. Ribeiro, J. d. E. B. Neto, C. Traina-Jr., and A. J. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method," Decision Support Systems **51,** 810–820 (2011), recent Advances in Data, Text, and Media Mining & Information Issues in Supply Chain and in Service System Design.

[84] R. S. Choraś, "Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems," International Journal of Biology and Biomedical Engineering **1,** 6–16 (2007).

[85] K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. M. Capretz, "Challenges for MapReduce in Big Data," In *Proceedings of the 2014 IEEE World Congress on Services*, SERVICES '14 pp. 182–189 (IEEE Computer Society, Washington, DC, USA, 2014).

[86] J. D. Bronzino, *The Biomedical Engineering Handbook* (CRC-Press, 2000), Vol. 2.

[87] D. Sankar and T. Thomas, "A New Fast Fractal Modeling Approach for the Detection of Microcalcifications in Mammograms," Journal of Digital Imaging **23,** 538–546 (2010).

[88] I. K. Maitra, S. Nag, and S. K. Bandyopadhyay, "Technique for Preprocessing of Digital Mammogram," Computer Methods and Programs in Biomedicine **107,** 175–188 (2012).

[89] S.Abinaya, R.Sivakumar, M.Karnan, D. Shankar, and M.Karthikeyan, "Detection of breast cancer in mammograms - a survey," International Journal of Advanced Research in Computer Science and Software Engineering **3,** 172–178 (2014).

[90] R. C. Gonzalez and R. E. Woods, *Digital Image Processing* (Addison-Wesley, 2002).

[91] B. Deshpande, H. Verma, and P. Deshpande, "Fuzzy Based Median Filtering for Removal of Salt-and-Pepper Noise," International Journal of Soft Computing and Engineering **2,** 76–80 (2012).

[92] J. Suckling *et al.*, "The mammographic image analysis society digital mammogram database," In *Proceedings of the 2nd International Workshop on Digital Mammography*, pp. 375–378 (1994).

[93] C.-H. Wei and C.-T. Li, "Calcification Descriptor and Relevance Feedback Learning Algorithms for Content-Based Mammogram Retrieval," In *Digital Mammography Lecture Notes in Computer Science*, **4046,** 307–314 (2006).

[94] P. Sathya and R. Kayalvizhi, "Optimum Multilevel Image Thresholding Based on Tsallis Entropy Method with Bacterial Foraging Algorithm," IJCSI International Journal of Computer Science Issues **7,** 336–343 (2010).

[95] R. Lederman, I. Leichter, E. Ratner, M. Abramov, A. Manevich, and J. Stoeckel, "Should CAD be used as a second reader? Exploring two alternative reading modes for CAD in screening mammography," In *10th international conference on Digital Mammography*, pp. 161–167 (2010).

[96] F. Janabi-Sharifi, in *Opto-Mechatronic Systems Handbook: Techniques and Applications*, *Handbook Series for Mechanical Engineering*, H. Cho, ed., (2002), Chap. 10.

[97] M. M. Mokji and S. A. R. Abu-Bakar, "Gray Level Co-Occurrence Matrix Computation Based On Haar Wavelet," In *4th International Conference on Computer Graphics, Imaging and Visualization (CGIV 2007), August 14-16, 2007, Bangkok, Thailand*, pp. 273–279 (2007).

[98] A. Papadopoulos, D. I. Fotiadis, and A. Likas, "An automatic microcalcification detection system based on a hybrid neural network classifier.," Artificial Intelligence in Medicine **25,** 149–167 (2002).

[99] S. Aksoy and R. M. Haralick, "Feature Normalization and Likelihood-based Similarity Measures for Image Retrieval," Pattern Recogn. Lett. **22,** 563–582 (2001).

[100] J. Collins and K. Okada, "A Comparative Study of Similarity Measures for Content-Based Medical Image Retrieval.," In *CLEF (Online Working Notes/Labs/Workshop)*, P. Forner, J. Karlgren, and C. Womser-Hacker, eds., (2012).

[101] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian Quadratic Discriminant Analysis.," Journal of Machine Learning Research **8,** 1277–1305 (2007).

[102] L. Ladha and T. Deepa, "Feature selection methods and algorithms," International Journal on Computer Science and Engineering (IJCSE) **3,** 1787–1797 (2011).

[103] Y. Peng, Z. Wu, and J. Jiang, "A Novel Feature Selection Approach for Biomedical Data Classification," Journal of Biomedical Informatics **43,** 15–23 (2010).

[104] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers.," Pattern Recognition **33,** 25–41 (2000).

[105] M. A. Alolfe, W. A. Mohamed, A.-B. M. Youssef, Y. M. Kadah, and A. S. Mohamed, "Feature Selection in Computer Aided Diagnostic System for Microcalcification Detection in Digital Mammograms," In *26th National Radio Science Conference'*, **26,** 1–9 (2009).

[106] J. F. D. Addison, S. Wermter, and G. Z. Arevian, "A comparison of feature extraction and selection techniques," In *Proceedings of the International Conference on Artificial Neural Networks*, pp. 212–215 (2003).

[107] R. Swiniarski and A. Swiniarska, "Comparison of Feature Extraction and Selection Methods in Mammogram Recognition," Annals of the New York Academy of Sciences **980,** 116–124 (2002).

[108] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," BMC Bioinformatics **7,** 1–16 (2006).

[109] S. M. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998).

[110] I. I. Andreadis, G. M. Spyrou, and K. Nikita, "A comparative study of image features for classification of breast microcalcifications," Measurement Science and Technology **22,** 1–9 (2011).

[111] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," International Journal of Engineering Research & Technology **1,** 1–6 (2012).

[112] L. B. Holder, I. Russell, Z. Markov, A. G. Pipe, and B. Carse, "Current And Future Trends In Feature Selection And Extraction For Classification Problems," International Journal of Pattern Recognition and Artificial Intelligence **19,** 133–142 (2005).

[113] M. Rizzi, M. D'Aloia, and B. Castagnolo, "Review: Health Care CAD Systems for Breast Microcalcification Cluster Detection," journal of medical and biological engineering **32,** 147–156 (2011).

[114] M. Mustra, M. Grgic, and K. Delac, "Enhancement of Microcalcifications in Digital Mammograms," In *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, **19,** 248–251 (2012).

[115] A. J. Alberg, J. W. Park, B. W. Hager, M. V. Brock, and M. Diener-West, "The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests," Journal of General Internal Medicine **19,** 460–465 (2004).

[116] C. E. Metz, "Basic principles of ROC analysis," Seminars in Nuclear Medicine **8,** 283–298 (1978).

[117] J. Beutel and M. Sonka, *Handbook of Medical Imaging: Medical image processing and analysis* (SPIE Press, 2000), Vol. 2.

[118] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations," In *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, pp. 1–9 (2010).

[119] F. N. Harirchi, P. Radparvar, H. Moghaddam, F. Dehghan, and M. Giti, "Two-Level Algorithm for MCs detection in mammograms using Diverse-Adaboost-SVM," In *20th International Conference on Pattern Recognition* , **20,** 269–272 (2010).

[120] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," Indian Journal of Ophthalmology **56,** 45–50 (2008).

[121] W.-V. Oh, K. Kim, Y.-J. Kim, H. Kang, J. Ro, and W. Moon, "Detection of Microcalcifications in Digital Mammograms Using Foveal Method," Journal of Korean Society of Medical Informatics **15,** 165–172 (2009).

[122] B. Mohanalin, P. Karla, and N. Kumar, "A novel automatic microcalcification detection technique using Tsallis entropy and a type II fuzzy index," Computers and Mathematics with Applications **60,** 2426–2432 (2010).

[123] S. Marcellin, D.-A. Zighed, and G. Ritschard, "Detection of breast cancer using an asymmetric entropy measure," In *Computational Statistics (COMPSTAT 06)*, A. Rizzi and M. Vichi, eds., Computational Statistics **XXV,** 975–982 (Springer, Heidelberg, Germany, 2006), on CD.

[124] B.Vijayalakshmi, R. Bhanumathi, and G. Suresh, "Study of Mammogram Micro calcification to Aid tumour detection using Artificial Neural Network Based Classifier," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering **3,** 644–650 (2014).

[125] F. Engelken, R. Bremme, U. Bick, S. Hammann-Kloss, and E. M. Fallenberg, "Factors affecting the rate of false positive marks in CAD in full-field digital mammography," European Journal of Radiology **81,** 844–848 (2012).