Pixel N-grams for Mammographic Image Classification

PRADNYA KULKARNI

Bachelor of Electronics Engineering (India)

Postgraduate Diploma in Information Technology
(University of Ballarat, Australia)

This thesis is submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Science and Technology
Federation University
PO Box 663
University Drive, Mount Helen
Ballarat, Victoria 3353
Australia

May 2017

Principal Supervisor: Associate Professor Andrew Stranieri
Associate Supervisor: Dr. Julien Ugon
Radiology Expert: Dr. Manish Mittal (LakeImaging Inc.)

Abstract

X-ray screening for breast cancer is an important public health initiative in the management of a leading cause of death for women. However, screening is expensive if mammograms are required to be manually assessed by radiologists. Moreover, manual screening is subject to perception and interpretation errors.

Computer aided detection/diagnosis (CAD) systems can help radiologists as computer algorithms are good at performing image analysis consistently and repetitively. However, image features that enhance CAD classification accuracies are necessary for CAD systems to be deployed. Many CAD systems have been developed but the specificity and sensitivity is not high; in part because of challenges inherent in identifying effective features to be initially extracted from raw images.

Existing feature extraction techniques can be grouped under three main approaches; statistical, spectral and structural. Statistical and spectral techniques provide global image features but often fail to distinguish between local pattern variations within an image. On the other hand, structural approach have given rise to the Bag-of-Visual-Words (BoVW) model, which captures local variations in an image, but typically do not consider spatial relationships between the visual "words". Moreover, statistical features and features based on BoVW models are computationally very expensive. Similarly, structural feature computation methods other than BoVW are also computationally expensive and strongly dependent upon algorithms that can segment an image to localize a region of interest likely to contain the tumour. Thus, classification algorithms using structural features require high resource computers. In order for a radiologist to classify the lesions on low resource computers such as Ipads, Tablets, and Mobile phones, in a remote location, it is necessary to develop computationally inexpensive classification algorithms.

Therefore, the overarching aim of this research is to discover a feature extraction/image representation model which can be used to classify mammographic lesions with high accuracy, sensitivity and specificity along with low computational cost. For this purpose a novel feature extraction technique called 'Pixel N-grams' is proposed. The Pixel N-grams approach is inspired from the character N-gram concept in text categorization. Here, N number of consecutive pixel intensities are considered in a particular direction. The image is then represented with the help of histogram of occurrences of the Pixel N-grams in an image.

Shape and texture of mammographic lesions play an important role in determining the malignancy of the lesion. It was hypothesized that the Pixel N-grams would be able to distinguish between various textures and shapes. Experiments carried out on benchmark texture databases and binary basic shapes database have demonstrated that the hypothesis was correct. Moreover, the Pixel N-grams were able to distinguish between various shapes irrespective of size and location of shape in an image.

The efficacy of the Pixel N-gram technique was tested on mammographic database of primary digital mammograms sourced from a radiological facility in Australia (LakeImaging Pty Ltd) and secondary digital mammograms (benchmark miniMIAS database). A senior radiologist from LakeImaging provided real time de-identified high resolution mammogram images with annotated regions of interests (which were used as groundtruth), and valuable radiological diagnostic knowledge. Two types of classifications were observed on these two datasets. Normal/abnormal classification useful for automated screening and circumscribed/speculation/normal classification useful for automated diagnosis of breast cancer. The classification results on both the mammography datasets using Pixel N-grams were promising.

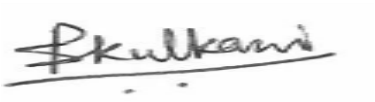
Classification performance (Fscore, sensitivity and specificity) using Pixel N-gram technique was observed to be significantly better than the existing techniques such as intensity histogram, co-occurrence matrix based features and comparable with the BoVW features. Further, Pixel N-gram features are found to be computationally less complex than the co-occurrence matrix based features as well as BoVW features paving the way for mammogram classification on low resource computers.

Although, the Pixel N-gram technique was designed for mammographic classification, it could be applied to other image classification applications such as diabetic retinopathy, histopathological image classification, lung tumour detection using CT images, brain tumour detection using MRI images, wound image classification and tooth decay classification using dentistry x-ray images. Further, texture and shape classification is also useful for classification of real world images outside the medical domain. Therefore, the pixel N-gram technique could be extended for applications such as classification of satellite imagery and other object detection tasks.

Statement of Originality

Except where explicit references are made, the text of this thesis contains no material published elsewhere or extracted in whole or in part from a thesis by which I have qualified for or have been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgement in the main text and bibliography of the thesis.

Pradnya Kulkarni



Signed:_____ Dated: 18 May 2017

List of Publications

1. Kulkarni, P., Stranieri, A., Ugon, J., Kulkarni, S., Mittal, M. (2017, April). Pixel N-grams for mammographic lesion classification, IEEE International Conference on Communication Systems, Computing and IT Applications, Mumbai, India, 55

2. Kulkarni, P., Stranieri, A., Ugon, J. Pixel N-grams. (2017) Size, Location and Resolution Invariance for Shape Classification, International Journal of Science Engineering and Management 1(8), 38-44

3. Kulkarni, P., Stranieri, A., Ugon, J., (2016, August). Texture Image Classification using Pixel N-grams, International Conference on Signal and Image Processing, Beijing, China, 137-141 **(Best Presentation Award)**

4. Kulkarni, P., Stranieri, A., Ugon, J., (2016, July). Pixel N-grams for mammograms classification, Annual Federation University Research Conference, Ballarat, Australia, 45

5. Kulkarni, P., Stranieri, A., Kulkarni, S. Ugon, J., & Mittal, M. (2015, July). Analysis and Comparison of Co-occurrence Matrix and Pixel N-gram Features for Mammographic Images, International Conference on Communication and Computing, Banglore, India, 7-14

6. Kulkarni, P., Stranieri, A. Kulkarni, S. (June 2014). A novel Architecture and Analysis of Challenges for Combining Text and Image for Medical Image Retrieval, International Journal of Infonomics, 7(1/2), 885-890

7. Kulkarni, P., Stranieri, A., Kulkarni, S. Ugon, J., & Mittal, M. (2014, April). Visual Character N-Grams for Classification and Retrieval of Radiological Images, International Journal of Multimedia and its Applications, 6(2), 35-49

8. Kulkarni, P., Stranieri, A., Kulkarni, S. Ugon, J., & Mittal, M. (2014, February). Hybrid Technique Based on N-gram and Neural Networks for Classification of Mammographic

Images, Second International Conference on Signal, Image Processing and Pattern Recognition, Sydney, Australia, 297–306

9. Kulkarni, P., Stranieri, A., Kulkarni, S., Mittal, M., & Ugon, J. (2013, November). The Integration of Image and Text Retrieval to Support Radiological Diagnosis, Annual UB Research conference, Ballarat, Australia, 25. **(Best Paper Award)**

Acknowledgement

I would like to take this golden opportunity to express my sincere thanks to my principle supervisor Assoc. Prof. Andrew Stranieri for his valuable guidance, understanding and tremendous support. He has been a constant inspiration throughout my research journey. I have been lucky to learn many things from him such as research skills, optimistic thinking, patience, hard work, writing skills and understanding psychology of other people around us. Without his support it would not have been possible for me to conquer the hill of doctoral level research work with change in country (huge change in the environment) and family commitments. I would also like to thank my associate supervisor Dr. Julien Ugon for guidance and support. Many thanks to senior radiologist Dr. Manish Mittal from LakeImaging, Ballarat for being actively involving in this PhD project, providing real time mammographic image data, medical knowledge and terminologies related to breast cancer and radiology.

I would like to extend my gratitude to my dearest husband Dr. Siddhivinayak Kulkarni who has been a major supporting pillar in this journey. The support from my young children Tejal and Tanmay has made this journey a pleasure and success. Encouragement and support from my father Bhalachandra Deshpande, mother Lata Deshpande, brother Mr. Prashant Deshpande, my in-laws Dr. Arvind Kulkarni, Anuradha Kulkarni and other friends and relatives has been outstanding.

I would also like to thank research services and Faculty of Science and Technology for providing all the necessary help and funding for completion of this PhD project. I am also thankful to our former head of school Prof. John Yearwood, current head of the school Dr. Jason Giri for prompt actions regarding any matter related to PhD work. Last but not the least I would like to thank few people (Dr. Philip Smith, Assoc. Prof. Peter Vamplew, Dr. Robert Watson, Assoc. Prof. Jim Sillitoe, Dr. Stephen Carey, Helen Wade, Rebecca Davis, Andrea Davis) who have helped and encouraged me a lot during this journey.

Statement of Ethics Approval

Approval has been granted to use existing data gathered by Lake Imaging.

Approval is for access to completely de-identified data supplied by Lake Imaging, who are responsible for gaining consent of the patients for use of their data.

Principal Researcher:	Andrew Stranieri	
Other/Student Researcher/s:	Pradnya Kulkarni	Manish Mittal Julien Ugon
School/Section:	SITE	
Project Number:	C13-011	
Project Title:	Pixel N-grams for mammographic image classification	
For the period:	27/8/2013 **to** 6/5/2016	

Please quote the Project No. in all correspondence regarding this application.

REPORTS TO HREC:

Annual reports for this project must be submitted to the Ethics Officer on:
27 August 2014
27 August 2015

A final report for this project must be submitted to the Ethics Officer on:
6 June 2016

These reports can be found at:
http://www.ballarat.edu.au/research/research-services/forms/ethics-forms

Ethics Officer
27 August 2013

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN – Artificial Neural Network

BI-RADS – Breast Imaging, Reporting and Data System

BoVW – Bag of Visual Words

BoVP – Bag of Visual Phrases

BoW – Bag of Words

CADe – Computer Aided Detection

CADx – Computer Aided Diagnosis

CBIR – Content Based Image Retrieval

CNN – Convolutional Neural Network

CPP – Curve Partitioning Points

DCT – Discrete Cosine Transform

DDSM – Digital Database for Screening Mammography

DICOM – Digital Imaging and Communications in Medicine

FP – False Positive

FN – False Negative

GET – Generic Edge Token

GLCM – Grey Level Co-occurrence Matrix

HGD – Histogram of Gradient Divergence

HOG – Histogram of Oriented Edges

LBP – Local Binary Pattern

LTP – Local Ternary Pattern

MIAS – Mammographic Image Analysis Society

MLP – Multi Layer Perceptron

MSWLD – Multiscale Spatial Weber Descriptor

NCL – Normalised

NLP – Natural Language Processing

PACS – Picture Archiving and Communication System

PCA – Principal Component Analysis

PCPG – Perceptual Curve Partitioning Points

PDA – Personal Digital Assistant

RBF – Radial Basis Function

ROI – Region of Interest

ROC – Receiver Operating Characteristics

RSNA – Radiological Society of North America

SGLDM – Spatial Grey Level Dependence Matrix

SIFT – Scale Invariant Feature Transform

SURF – Speeded Up Robust Transform

SVM – Support Vector Machine

TP – True Positive

TN – True Negative

UIUC – University of Illinois at Urbana Champaign

WBCD – Wisconsin Breast Cancer Dataset

WHC – World Health Care

WHO – World Health Organisation

1 Introduction

This chapter provides some background information about breast cancer, screening mammography and computer aided detection/diagnosis. The motivation behind this project is explained with a research gap introducing the novel Pixel N-gram model for image representation. The research questions to be answered are mentioned followed by the significant contributions of this project. Lastly, the organisation of this thesis is detailed followed by a brief chapter summary.

The main goal of this research is to discover a new feature extraction technique for mammographic lesion classification. Shape and texture of mammographic lesions play an important role in classification of the lesions. It is hypothesized that the novel Pixel N-grams features are able to distinguish between various textures and shapes and are therefore useful and efficient for mammographic lesion classification.

In this work classification performance is measured with the help of evaluation measures such as accuracy, sensitivity, specificity, precision and Fscore. Classification accuracy is the number of correct predictions out of total predictions made. Other than the Accuracy, Sensitivity and Specificity are the two very important measures for detection of certain disease from clinical perspective. The higher numerical value of sensitivity suggests lower number of false-positive results. A test with high sensitivity tends to capture all possible positive conditions without missing anyone. Specificity measures the proportion of negatives that are correctly classified. Thus, high specificity is better for ruling out a particular disease condition. Precision tells how many of the positively classified instances were relevant. Fscore is a harmonic mean of precision and recall/sensitivity. Receiver Operating Characteristic (ROC) curve is plot of sensitivity/true positive rate on X axis and (1-specificity/true negative rate) on Y axis. Thus the trade-off between false negative and false positive errors can be achieved by analysing the ROC curve.

1.1 Background and Motivation

1.1.1 Breast Cancer

Breast cancer is the most frequently diagnosed cancer among women (140 out of 184 countries) and a common cause of death (Ferlay et al., 2010). About 522,000 deaths were reported due to breast cancer in 2012 (World Health Organisation, 2013). According to the (World Health

Organisation (WHO)), breast cancer has increased by more than 20% since 2008. About 1.7 million women were diagnosed with breast cancer in 2012.

Breast cancer is the unrestrained growth of the breast cells. There are two main types of abnormalities found in the breast: tumours and calcifications. A tumour is a mass of abnormal tissue. Two types of tumours exist; one type is non-cancerous is benign and the other is cancerous which damages the surrounding tissue and is malignant. On the other hand, small calcium deposits developed in women's breast tissue are called breast calcifications. These are usually benign, however, certain types of calcifications (micro-calcifications) are early signs of breast cancer. Early detection of breast cancer involves finding and diagnosing lesions before symptoms are obvious. Screening refers to the tests and their interpretations for early detection of breast cancer. Screening programs include examining women without any symptoms, thus increasing the healthcare costs. Also, screening may produce false positive results. The false positives tend to increase women's anxiety and unnecessary biopsies and have an impact on the psychological state of women under examination (Bleyer & Welch, 2012). However, screening detects cancer at an early stage when it is most likely to be confined to the breast and is treatable (National Breast Cancer Foundation, 2016). Breast screening can thus save thousands of lives each year. Screening for breast cancer is explained in the Section 1.1.2 below.

1.1.2 Screening Mammography

Mammography is a type of breast imaging where low dose x-rays are used. During mammography, the breast is compressed while an x-ray source emits radiation from one side to be captured on the other side of breast with the help of film or an electronic device. Different tissues appear as different levels of grey on mammographic image. Conventional mammograms were film based, however, digital mammograms are more commonly used now. There are two types of digital mammograms. X-ray beams directly recorded as digital images are called primary digital mammograms[1] whereas film based mammograms scanned into digital format are known as secondary digital mammograms[2].

[1] Truly digital mammograms/ primary digital mammograms are digital mammograms directly generated with the help of advanced imaging equipment.

[2] Secondary digital mammograms are conventional film based mammograms digitised with the help of a scanner.

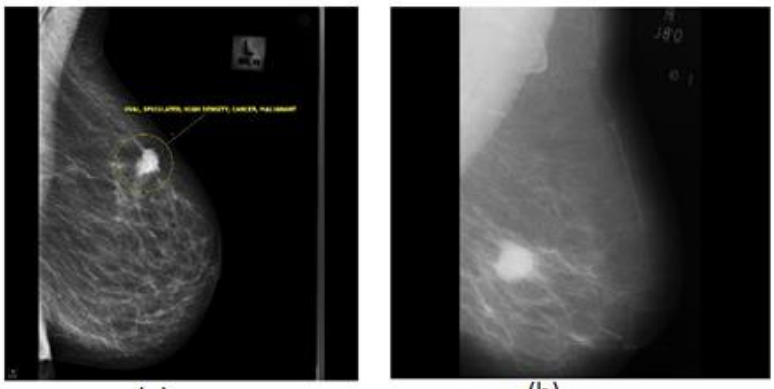


Figure 1.1 Sample Primary and Secondary Digital Mammograms

Figure 1.1(a) shows a sample primary digital mammogram from LakeImaging dataset. LakeImaging Pvt. Ltd. Inc. (LakeImaging, 2017) is a large diagnostic radiology provider in Australia. Experienced radiologists from LakeImaging have specially prepared the dataset of mammograms for this project. The images were mammograms collected from various patients during 2013 and 2014; manually annotated providing information such as shape, margins of lesions, background tissue density, type of abnormality and diagnosis related to the case. On the other hand Figure 1.1(b) shows a sample secondary digital mammogram from the miniMIAS dataset (Suckling et al., 1994). The miniMIAS database is a benchmark database provided by the Mammographic Image Analysis Society for research purposes. Film based mammograms are digitized with the help of a scanner and clipped or padded so as to get equal resolution of 1024 × 1024 pixels for generating the miniMIAS database. Digital mammograms are more convenient and inexpensive than film based mammograms as computers can be deployed to present, search, process and transmit the images quickly.

Diagnostic mammography refers to the mammography used for diagnosing breast diseases in women experiencing symptoms such as pain, lump or nipple discharge. On the other hand screening mammography is used for early detection of breast cancer in women experiencing no symptoms. Screening mammography is the most reliable and proven method for early detection of breast cancer (Suleiman, Lewis, Georgian-Smith, Evanoff, & McEntee, 2014). This is possible because mammography can show changes in the breast up to two years before a patient or physician can feel them ((RSNA), 2016). When a cancer is detected at an early stage it is treatable giving the patient higher chances of survival. It has been noted that screening mammography reduces the mortality rate by about 15 to 25% (Løberg, Lousdal, Bretthauer, & Kalager, 2015).

Although, screening mammography is the most reliable method for early detection of breast cancer, interpretation of lesions in mammograms is a very difficult and time consuming task for radiologists as the features of the abnormality are obscured or can be similar to those of the normal tissue (Sickles et al., 2005). It has been observed that radiologists usually struggle to maintain high interpretation accuracy while trying to fulfil the productivity targets (Akgül et al., 2011). Also, the breast cancer screening sensitivity of radiologists is only about 68% and specificity is about 75% (H. Cheng et al., 2006; Newton, 2016).

A radiologist is prone to make two types of errors while diagnosing breast diseases using mammograms : 1) perception/reading errors where a radiologist is unable to notice the abnormality due to fatigue and stress depending on time of the day and his/her health conditions (Palazzetti et al., 2016; Pow, Mello-Thoms, & Brennan, 2016), and 2) interpretation error where a radiologist notices the abnormality, but fails to interpret it correctly due to the lack of experience (Muramatsu et al., 2005). The common reasons for perception errors are lesions at the edge of the breast image (high contrast), lesions that are obscured due to overlying breast tissue and mass lesions on highly dense background tissue. On the other hand there are many reasons for occurrence of interpretation errors. Firstly, the lesion may be very small to prompt an action. Secondly, lucent areas within a dense background tissue may be perceived as fat and may be considered to be benign. A third reason is that the lesion may look similar to variations in appearance of normal breast tissue (Birdwell, Bandodkar, & Ikeda, 2005). For example, an inexperienced radiologist may get confused between a mass and a cyst (a small harmless sac filled with fluid which feels like lump) as both can look very similar on mammographic images. Perception and interpretation errors can again be divided into two categories namely false-positives and false-negatives. A false-positive occurs when a radiologist identifies an area as cancerous when it is actually non-cancerous. On the other hand if the radiologist fails to detect the abnormality it is called false-negative. False positives are not fatal but can increase patient's anxiety and health-care costs by leading to unnecessary treatments whereas false-negatives can be very serious and can reduce the survival chances of a women by delaying diagnosis and treatment (Metsälä, Pajukari, & Aro, 2012). Thus, it is necessary to develop accurate and affordable approach to early detection of breast cancer all over the world.

Further, the number of screening images generated every day are enormous. For example a typical large breast screening unit in the UK National Health Service Breast Screening

Programme was responsible for screening as many as 38,000 women each year producing approximately 1000 mammograms per week (Boggis & Astley, 2000). Hence there has been a concomitant need for computerized tools in the diagnostic process to help radiologists. Computer aided detection and diagnosis is discussed next.

1.1.3 Computer Aided Detection/Diagnosis (CAD)

Computer systems that assist radiologists in the interpretation of medical images are known as CAD systems and they mainly fall into two categories. First is CADe (Computer Aided Detection) which identify suspicious regions in the image, and the second is CADx (Computer Aided Diagnosis) that classifies the suspicious regions.

CAD systems help in minimizing perception as well as interpretation errors. This is because computer programs that process images can perform certain tasks repetitively and consistently. Although, a computer program may never be able to achieve the level of knowledge and cognitive capability of a radiologist, it can be better in recognising abnormality patterns in an image. Also, automated classification algorithms can help radiologists by providing a second opinion about the interpretation of lesions. Moreover, as these programs execute quickly they can help radiologists achieve productivity goals.

Some CAD systems tend to overstress the sensitivity[3] which comes at the expense of specificity[4]. This results in increase in false positive rate increasing unnecessary biopsies. More information about CAD systems can be found in reviews by (Jalalian et al., 2013; Rangayyan, Ayres, & Desautels, 2007).

Discussion with senior radiologists at LakeImaging Pvt. Ltd (diagnostic radiology provider in central and western Victoria) revealed that existing CAD systems are slower and still need to be improved with respect to accuracy, specificity and computational time. LakeImaging provided a database of mammograms where abnormalities were marked and annotated. Also, LakeImaging radiologist, Dr Manish Mittal, providing the clinical knowledge necessary to interpret the mammograms and provide insight into actual radiology workflow for screening as well as diagnosis of breast cancer. Further, they also indicated that it would be convenient to have CAD systems run on mobile devices such as smartphones.

[3] Sensitivity measures the proportion of positives correctly identified.
[4] Specificity measures the proportion of negatives correctly identified

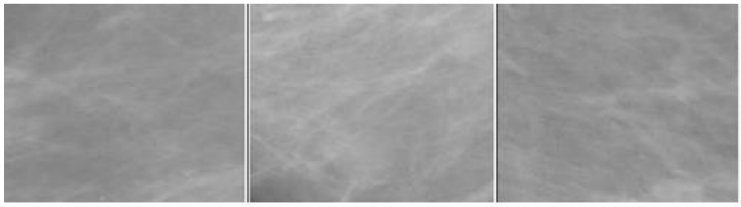
CAD systems have been used for detecting calcifications and mass/tumours in recent years (Tang, Rangayyan, Xu, El Naqa, & Yang, 2009). Mass detection is more difficult than detection of calcifications. Basically, mass lesions have three important characteristics namely texture, shape and margins (Wei, Li, & Huang, 2011). The features explaining the aforementioned characteristics determine the degree of likelihood of the lesion being benign or malignant. For example irregular shapes with speculated margins are most likely to be malignant; whereas, lesions with oval, round or lobular shape with smooth margins are likely to be benign (Varela, Timp, & Karssemeijer, 2006). Thus, the breast cancer detection can be thought of as an image classification problem based on the shape, texture and margin/boundary features. Various features used for computer aided diagnosis of breast cancer using screening mammography are explained in the Section 1.1.4 ahead.

1.1.4 Features used for classification of mammographic lesions

Computer aided diagnosis (CADx) involves classification of mammographic regions of interest (ROI). The ROI is the region occupied by the abnormality/lesion on a mammographic image. Figure 1.2 shows some examples of ROIs containing various types of lesions such as circumscribed masses, speculated masses and normal breast tissue.

(a) Circumscribed masses

(b) Speculated masses



(c) Normal breast region

Figure 1.2 Example lesions in mammograms (miniMIAS dataset)

As described earlier, texture and shape are the two important characteristics of mammographic lesions. Early work in mammographic classification based on texture includes systems based on various main texture features such as intensity histogram, Grey Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), Gabor filters and wavelet transforms. Similarly, widely used shape features include curvature functions, Zernike moments, Fourier transform, polynomial approximation, and Bounding boxes. Various texture and shape features used for mammographic classification will be discussed in Chapter 2.

All of the above mentioned features describe the image globally (as a whole) and are known as global image features. However, global features are not useful for classification or retrieval of certain image domain. For example, in medical radiology clinically useful information such as lesion/tumours are characterised by different tissue densities which appear as grey level variations in highly localized regions of an image. Here the number of pathology bearing pixels are relatively small and thus global image features will not be sufficient for diagnostic classification or retrieval (Shyu et al., 1998). Further, the pathology regions in medical images often do not possess sharp edges and contours. Therefore, automatic segmentation algorithms do not work accurately on these images (Shyu et al., 1998).

To capture local variations in an image the Bag-of-Visual-Words (BoVW) model has emerged recently. This model is originated from the Bag-of-Words (BoW) model from the text categorization domain. For example a document containing the sentences "Meera likes apples. Meera also likes oranges. Apples are sweet" can be represented with the help of number of occurrences of words in the generated vocabulary "Meera", "likes", "apples", "oranges", "sweet", "also", "are". Similarly, in the BoVW approach images are represented with the help of frequency of occurrence of the visual words (local patch features) present in an image. Recent studies on image classification have shown that the use of Bag-of-Visual-Words (BoVW) model (Y. Li, Chen, Rohde, Yao, & Cheng, 2015) for image representation provides better classification results than the existing low level global features such as colour, texture and shape mentioned above (Pedrosa & Traina, 2013).

1.1.5 N-grams

Even though the BoVW model has been successful in the image classification domain, this approach has some limitations. The limitations of this approach are: firstly spatial relationships between the visual words are not considered; secondly, the algorithms are computationally complex and thirdly, they are prone to noisy words. Hence, it becomes a priority to exploit an

image representation model using which the aforementioned BoVW limitations could be overcome.

Spatial relationships can be considered by taking two analogies from the text retrieval context. There are two main concepts used in the text retrieval context, namely: word N-grams and character N-grams. A word N-gram is a phrase formed by N consecutive words in a document. The use of word N-grams is an efficient approach in text retrieval (Suen, 1979). Similarly, instead of taking the image as a set of isolated visual words, consideration of visual phrases formed by N consecutive visual words would lead to more meaningful image representation. In case of an image it is achieved by using visual dictionaries and the idea is to consider visual patterns similar to textual words.

The use of word N-grams significantly improves the retrieval precision as well as classification accuracy with respect to the BoVW approach (Pedrosa & Traina, 2013). However, there still exist few drawbacks. The N-gram dictionary size and hence computational complexity goes on increasing as N is increased. Moreover, some of the information is lost during quantization step and noisy phrases are created due to noisy visual words.

On the other hand, character N-grams are sequences of N consecutive letters in a sentence. For example, the 3-grams in the phrase "this dog" are "thi, his, is_, s_d, _do, dog"; the four grams are "this, his_, is_d, s_do, _dog". Character N-gram model has worked exceptionally well in text classification than word N-gram model for languages such as Chinese, which lack specific word boundaries (Kanaris, Kanaris, Houvardas, & Stamatatos, 2007). In text categorization domain, it is evident that character N-gram algorithms are language independent, simple and computationally less expensive, yet provide comparable or better results than word N-grams (Kanaris et al., 2007).

Motivated by character N-grams in text a novel image representation model called **'Pixel N-grams'** is proposed in this PhD project. To the best of our knowledge this idea has not been explored for image classification application. Intensity values of N consecutive pixels in an image are considered in this approach and the image can be represented with the help of frequency of occurrence of various Pixel N-grams present in an image. The advantages of the Pixel N-gram approach over BoVW and word N-grams approach are computational cost effectiveness, spatial relationship consideration and no loss of information.

1.1.6 Focus of the project and other possible extensions

It is clear that the screening mammograms is a time consuming and demanding job for the radiologists and the sensitivity is affected by the factors such as fatigue, workload and experience level of the radiologist. The chances of missing a cancer lesion can be improved by double reading from another radiologist. However, double reading increases the workload burden on the radiologists. The workload burden can be reduced by using CAD systems (Tang et al., 2009). Existing CAD systems can be improved in terms of the accuracy, sensitivity as well as specificity. Further, if these systems are made computationally efficient, they could help in improving the turnaround time of screening process devices such as Personal Digital Assistant (PDA), Mobile Phones and IPad, provide convenience for the radiologist and can be used for evaluation of digital radiological images, at least, in emergency events or for preliminary diagnosis (I Drnasin, Gogić, & Drnasin, 2010). Using these devices, a radiologist can connect to the PACS server[5] from anywhere, anytime in the world to view, diagnose or report on a particular cases. This facility is quite helpful for medical teleconsultation in a wireless covered hospital environment. Further, the radiology images (PACS) are big in size and therefore require good memory size and high processor speed requirement. The main challenge for developing applications for mobile devices is their low memory, processing speed and display quality (Ivan Drnasin & Grgic, 2010). It is clear that if the CAD algorithms are designed to be computationally cost effective then they would be helpful for use with mobile devices.

This PhD project is mainly focused on developing a new feature extraction method for classification of mammograms in order to improve the classification accuracy, sensitivity as well as specificity. At the same time an effort is made to design the feature extraction algorithm to be computationally cost effective. A novel Pixel N-grams image representation model for mammograms classification/breast cancer diagnosis inspired from character N-gram concept in text categorization is proposed. Further, texture as well as shape classification using Pixel N-gram features is also explored as mammographic lesions are mainly characterised by texture and shape. Mammographic image classification for primary as well as secondary digital images is also analysed.

[5] Picture Archiving and Communications System

However, the Pixel N-gram image representation is not limited to mammographic images and can be applied to other image classification applications such as diabetic retinopathy, lung cancer detection, brain tumour detection.

1.2 Research Questions

The primary objective of the current research is to discover a new feature extraction technique for mammographic lesion classification. The main research question to meet the objective of the project is:

How can Pixel N-gram features be used for effective classification of mammographic lesions?

In order to address the main research question, the following sub-questions are investigated.

- Can character N-gram approach in text categorization context be successfully used for image classification?
- How well can the classifiers trained with Pixel N-gram features distinguish between various textures?
- How well can classifiers trained with Pixel N-gram features distinguish between various shapes?
- Is the classification using Pixel N-gram features independent of size and location of the shape?
- How effective are the Pixel N-gram features in classifying between abnormal and normal mammographic images?
- How effective are the Pixel N-gram features in classifying between circumscribed, speculated and normal mammographic lesions?
- Is the generation of Pixel N-grams features computationally less expensive than the generation of other features?

1.3 Intellectual Contributions and Significance

The major contributions of this PhD project are:

- One of the main contributions of this research is a simple and computationally inexpensive algorithm for image feature extraction. A novel feature extraction and image representation model called Pixel N-gram model is advanced.

- A second contribution is the analysis of Pixel N-gram model for texture image classification.
- A third contribution is to explore the Pixel N-gram model for shape classification. Also, analysing to what extent the shape classification using Pixel N-gram features is size and location invariant.
- A fourth contribution is to exploit Pixel N-gram model for classification of mammographic lesions using primary (images generated with full field digital mammography unit) as well as secondary digital mammograms (film based mammograms digitized with the help of scanner).
- A fifth contribution is to access the computational complexity of Pixel N-gram model in comparison with other feature extraction techniques.

1.4 Organisation of Thesis

The thesis is organised into six chapters as follows.

The current chapter provides background information related to breast cancer, its early detection and need for computer aided detection/diagnosis. It clarifies the motivation behind the research undertaken. It also lists the research questions and significant contribution of this PhD project.

Chapter 2 provides an extensive review of the related work in the mammographic lesion classification, CAD systems, various feature extraction techniques and N-gram model for image representation.

Chapter 3 details the methodology used for mammographic as well as texture and shape image classification, including datasets used, pre-processing required, experimental procedures and evaluation measures.

The experimental results on texture and shape dataset are noted in Chapter 4. The analysis of the results and observations on these texture and shape dataset are also discussed in this chapter.

The experimental results on the primary digital mammograms (LakeImaging dataset) and secondary digital mammograms (miniMIAS) are detailed and analysed in Chapter 5.

Finally, Chapter 6 concludes the thesis. The limitations of the work presented in this thesis and the directions for future research are discussed in this chapter.

1.5 Chapter Summary

In this chapter, breast cancer is introduced and the importance of early detection of breast cancer is explained. The most reliable method for early detection of breast cancer is screening mammography. The challenges inherent in interpretation of the mammograms are outlined. Further, the need for computer aided detection/diagnosis in order to help the radiologist to increase the interpretation accuracy, sensitivity and reduce the workload is detailed. CAD systems mainly classify lesions into normal and abnormal regions and the image features play an important role in the classification accuracy. The N-gram model for image representation, inspired from text categorization is introduced. Motivated from the N-gram model a novel image representation model is proposed in this PhD work. The research questions and the contributions of this research are then listed followed by the organisation of the thesis. The next chapter (literature review) provides a summary of the related work in this area.

2 Literature Review

2.1 Digital Mammography and Breast Cancer Detection and Diagnosis

Breast cancer is the most widely detected cancer in women all over the world (World Cancer Research Fund International, 2017). Early detection of breast cancer opens up various treatment options thereby increasing the survival chances of a patient (Jalalian et al., 2013). Mammography (breast x-rays) screening is a proven method for early detection of breast cancer. However, interpretation of mammographic images is a tedious and difficult task for the radiologist due to complexities like variability in appearance of lesions and density of the surrounding tissue (McKenzie, 2014). In addition with enormous number of mammograms generated every day, radiologists are under tremendous workload pressure. Therefore, they are prone to make errors in diagnosis simply due to tiredness, fatigue or insufficient experience levels (Evans, Birdwell, & Wolfe, 2013; te Brake, Karssemeijer, & Hendriks, 1998).

In recent years, computer aided systems are deployed to help radiologists perform accurate and efficient detection and diagnoses for breast cancer. This is plausible as computers can perform image analysis (Eberl, Fox, Edge, Carter, & Mahoney, 2006) consistently and repetitively. Computer aided detection/diagnosis (CAD) systems are mainly grouped into two categories namely CADe (computer aided detection) and CADx (computer aided diagnosis) (Jalalian et al., 2013). CADe systems involve classifying regions on mammographic image into two categories namely abnormal (suspicious) and normal. Thus CADe systems try to reduce perception errors by prompting the suspicious regions to the radiologist. On the other hand, CADx systems are required to give extra information to the radiologists such as shape of the lesion, margin of the lesion and cancer severity. CADx systems involve classification of abnormal regions into various categories in order to help radiologist diagnose the abnormality in an efficient and accurate way. One of the classifications in CADx system is classification of abnormal regions into benign or malignant (cancerous). Another classification involved in CADx system is fine-grained classification into categories such as circumscribed lesions, speculation lesions, calcifications and normal. Further, the CADx systems can also involve classifying the case into the BI-RADS shape classes such as round, oval, lobular and irregular classes (Vadivel & Surendiran, 2013). Breast cancer cases can further be classified according to the severity of cancer or assessment (Eberl et al., 2006).

Apart from detection and diagnosis of breast cancer, the classification of mammographic lesions can be very useful for retrieving images for training inexperienced radiologists. Mammographic image interpretation consists of three tasks; perception of image findings, interpretation of these findings to render a differential diagnosis and recommendations for clinical management (Muramatsu et al., 2005). The perception of image findings and their interpretation requires two types of knowledge. One is formalized domain knowledge which is present in texts and other written documentation. Second is implicit or tacit knowledge consisting of radiologists' individual expertise, organisational practices and past cases (Montani and Bellazzi, 2002). For less experienced radiologists this practical data of past cases is not available in his/her memory. Therefore, a facility to search for past similar cases can be beneficial in diagnostic decision making.

The presentation of similar images has shown to increase the confidence level of the radiologist (Muramatsu, Schmidt, Shiraishi, Li, & Doi, 2010) while interpreting a new image and also to improve radiologists' performance (Horsch et al., 2006; Nakayama, Abe, Shiraishi, & Doi, 2009). The classification of mammographic images can help speed up the retrieval of similar cases and thus, is helpful for training less experienced radiologists. Further, the classification of mammograms is also useful for medical research. For instance if a researcher wants to see the cases of breast cancer in stage1 in the last 5 years and analyse the images, mammographic classification can automate the assembly of the required dataset.

CADe or computer aided detection systems are designed to detect and locate suspicious regions (Birdwell et al., 2005; Freer & Ulissey, 2001). The suspicious regions or ROIs are the regions containing abnormalities. Two types of methods are common for CADe systems. Pixel based method and region based method (H.-D. Li, Kallergi, Clarke, Jain, & Clark, 1995). Pixel based methods features are computed from every pixel in an image and then the classification into normal or suspicious pixels is carried out. Later the pixels classified as suspicious are grouped together to form suspicious regions. In region based methods the ROIs are extracted using a segmentation or a filtering approach before computing the features from the ROI. The classification into normal or suspicious area is then performed by using the features of the ROI. Further, as the ROI features correlate with the important diagnostic information such as shape and margin of the lesion they are much better than the pixel based methods. However, the accuracy of region based methods is heavily dependent upon the accuracy of the segmentation

algorithm which could be very low in case of dense background tissues (Buist, Porter, Lehman, Taplin, & White, 2004).

On the other hand CADx systems are designed to help the radiologists in diagnosing suspicious lesions. The most frequent indications of breast cancer on mammograms are masses and microcalcifications (Sampat, Markey, & Bovik, 2005). Masses are group of cells having higher density than the surrounding tissues. Microcalcifications are small calcium deposits appearing as opacities in mammograms. Radiologists use shape and margin properties of the mass to judge the malignancy of the mass (Halls, 2017). Masses with speculated margins are most likely to be malignant however, not all malignant masses are speculated. Small, clustered and irregular shaped calcifications (microcalcifications) are most likely to be malignant whereas rounded, large and regular shaped diffusely distributed calcifications (macrocalcifications) are usually benign. Thus, malignancy of calcifications depends on the size, shape and their distribution. An exhaustive review for detection and classification of microcalcifications can be found in (H.-D. Cheng, Cai, Chen, Hu, & Lou, 2003; Lewenstein & Urbaniak, 2016).

In general, it is evident that the likelihood of malignancy of a lesion is dependent upon the shape, texture/density and margins of a lesion (Wei et al., 2011). Therefore, CADx systems are designed to extract the features which can model these important characteristics. However, the development of CADx systems has been challenging for the following reasons. Firstly, very high accuracy is needed as misclassifications can have serious effects on patient care. Secondly, the abnormalities can be occluded by dense tissue. Finally, the appearance of the abnormalities may differ drastically making it difficult to analyse (Jirari, 2008). More information about computer aided detection/diagnosis of breast cancer can be found in the review articles (Doi, 2009; Jalalian et al., 2013; Mohanty, Champati, Swain, & Lenka, 2011; Rangayyan et al., 2007; Sampat, Markey, et al., 2005).

Many commercial CAD systems have been implemented. Some examples of commercial CAD systems are Imagechecker designed by R2 Technology (Hologic Inc), M-Reader (M-reader, 2001), Second Look implemented by CADx Inc. (iCAD, 2000), Mammex Tr produced by Scanis Inc.(Scanis Inc), and ImageClear developed by Titan Systems Corp.'s. A study of the effect of CAD on screening observed that the number of cancers detected increased by 19.5%, and the early-stage malignancies detected increased from 73% to 78% (Freer & Ulissey, 2001) following the introduction of CAD. Further, It has been reported that CAD has provided a

gain of approximately 10-20% in the early detection of breast cancers on mammograms (Doi, 2009).

Despite many efforts in designing CAD systems, high sensitivity comes at the cost of low specificity resulting in many false positives (Jalalian et al., 2013). Considering the emotional stress for patients and higher healthcare costs due to unnecessary biopsies, there is a great need to improve CAD systems in terms of achieving high accuracy, specificity and sensitivity. Further, the area of CADx systems is not well explored and has the potential to improve the radiologist's diagnostic accuracy and training of less experienced radiologists. Also, in CADe, CADx, and retrieval of cases for training/research purposes, faster feature extraction algorithms can greatly increase efficiency and save a huge amount of time for the radiologists, releasing the workload burden to a fair extent. Hence, the development of faster and accurate image processing algorithms for automated detection/diagnosis/classification is needed for the future.

One of the most important steps in accurate classification of images is the extraction of features which can be used by classification algorithms to distinguish between various image classes. Many different features have been advanced for the classification of mammographic images. These features are described in the Section 2.2.

2.2 Features for Image Classification

In general image features can be grouped under three main categories: 1) primitive/low level features (colour, texture, shape, spatial location) 2) midlevel features (objects for example human, car, tumour) and, 3) high level features (scenes)(Eakins & Graham, 1999). Mid and high level features are called semantic features. High level features/semantic features are used by humans to describe an image such as an object, scene or action. However, most of the existing classification work involves the use of primitive features mainly because the semantic features (identifying objects or scenes) are very difficult to extract.

Mammograms are mostly grey scale images and hence colour information is not available. Various texture and shape features used for classification of mammograms are explained in Section 2.2.1 and Section 2.2.2 respectively.

2.2.1 Texture Features

Texture is defined as a quantitative measure of the arrangement of intensities in an image. As depicted earlier, texture is an important characteristic of lesions including tumours, cysts and calcifications in mammograms. Texture can be modelled using various approaches namely statistical, structural and spectral, discussed next.

2.2.1.1 Statistical approach

Statistical approaches represent texture using non-deterministic properties capturing the distributions and associations between the grey levels in an image. Statistical features are known to distinguish between various textures very well. Many researchers have tried various statistical techniques for classifying mammographic lesions.

First order statistical features are used traditionally for texture classification. An Intensity histogram is one of the commonly used, simple statistical approach for modelling texture. An intensity histogram shows frequency of occurrence of intensity levels in an image. Intensity histogram is thus a concise summary of the statistical information in an image. Probability density of occurrence of the intensity in an image can be calculated by dividing the histogram counts by total number of pixels in an image. Different images can have the same intensity histogram. First order statistical features such as mean, variance, skewness, kurtosis are thus computed from the histogram and are called central moments. Statistical features based on histogram (mean, standard deviation, smoothness, third moment, uniformity and entropy) were used for classifying breast tissue into four categories namely fatty, uncompressed fatty, dense and high density (Sheshadri & Kandaswamy, 2006). Classification accuracy by testing this approach on miniMIAS database (Islam, Ahmadi, & Sid-Ahmed, 2010) was 78%. Statistical features such as mean, standard deviation, smoothness, entropy, skewness, kurtosis and uniformity was used for classification of ROIs into benign and malignant categories. A three-layer artificial neural network trained using the above mentioned 7 features provided a sensitivity of 90.91% and specificity of 83.87% which was significantly higher than the radiologists average sensitivity (75%). The computational time recorded was about 15-20 msec for each classification. The experiments were carried out on 69 images of miniMIAS dataset (Islam et al., 2010).

Classification into mass and normal tissue has been achieved using contrast, intensity and location (te Brake, Karssemeijer, & Hendriks, 2000). The method was tested on 132 mammograms from Nijmegen screening program and 772 mammograms from DDSM database. It was observed that these features were successful in discriminating masses from false positive detections. Thus an accuracy of 75% was achieved with a specificity level of 0.1 false positive per image. First order statistical features are computationally simple and quick to calculate. The first-order statistical features provide information related to the grey level distribution of the image. However, they do not provide any information about the relative positions of the various grey levels within the image. These features will not be able to measure whether all low-value grey levels are positioned together, or they are interchanged with the high-value grey levels. Thus first order statistical features are unable to provide information about the spatial relationships among the grey level pixels(Aggarwal & Agrawal, 2012).

Thus in order to model the spatial information between the neighbouring pixels the second order statistical features were developed. Further, a study of human texture discrimination in terms of texture statistical properties indicates that the textures in grey scale images are distinguished spontaneously only if they differ in second order moments (Julesz, 1975). Second order histogram is also known as Grey Level Co-occurrence matrix (GLCM) (Haralick, Shanmugam, & Dinstein, 1973) and is a statistical method of texture description based on repeated occurrence of grey-level configuration in the texture. Grey level co-occurrence matrix is a matrix of relative frequencies P(i,j) describing how frequently two pixels with grey-levels i and j appear at a distance d in the direction θ as shown in Figure 2.1 Computation of grey level co-occurrence matrix can be calculated in several directions and distances.

Haralick (Haralick et al., 1973) proposed 14 second order statistical features extracted from the co-occurrence matrix in order to calculate the similarity between two images. The 14 second order statistical features proposed by him were: Angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, entropy, sum entropy, difference variance, difference entropy, information measure of correlation 1, information measure of correlation 2 and maximum correlation coefficient and are computed using the GLCM.

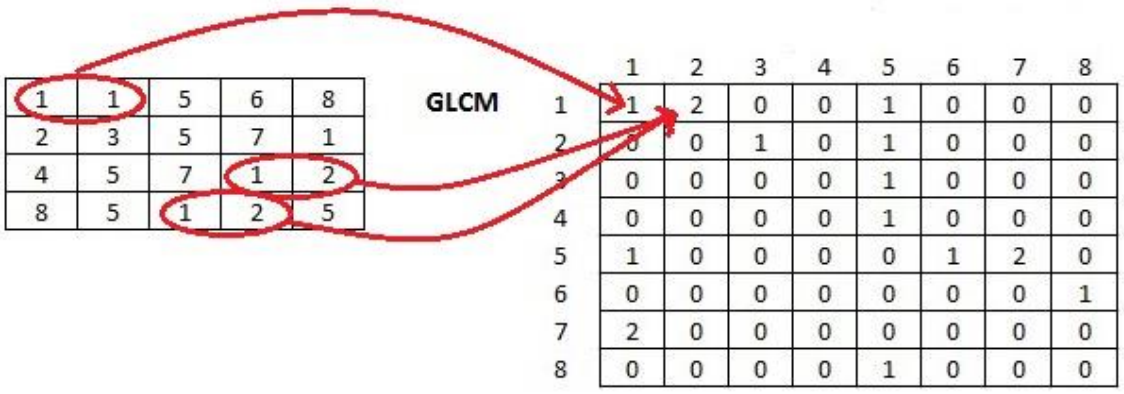


Figure 2.1 Computation of grey level co-occurrence matrix

The equations for computing these Haralick's features are given in the Table 2.1.

$$Pij = (i,j)th\ entry\ in\ a\ normalized\ gray\ \ level\ co-occurrence\ \ matrix$$

$$P_x(i) = \ ith\ entry\ in\ the\ marginal\ probability\ matrix\ obtained\ by\ summing\ the\ rows\ of\ P(i,j)$$

$$N_g = \ Number\ of\ distinct\ grey\ levels\ in\ the\ quantized\ image$$

$$P_y(j) = \sum_{i=1}^{N_g} P(i,j)$$

$$P_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j), \qquad k = 2,3,\dots 2N_g, \qquad i+j=k$$

$$P_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j), \qquad k = 0,1,\dots N_g-1, \qquad |i-j|=k$$

$$HXY = -\sum_i\sum_j P(i,j)\log P(i,j)$$

$$HXY1 = -\sum_i\sum_j p(i,j)\log[P_x(i)P_y(j)]$$

$$HXY2 = -\sum_i\sum_j P_x(i)P_y(j)\log[P_x(i)P_y(j)]$$

Table 2.1 Haralick's features (co-occurrence matrix)

Feature	Formula
Angular Second Moment (Energy)	$\sum_i \sum_j (P_{ij})^2$
Contrast	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-j)^2 P_{ij}$
Correlation	$\sum_i \sum_j \frac{(ij)P_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y}$ Where $\mu_x, \mu_y, \sigma_x \; and \; \sigma_y$ are means and standard deviations of P_x and P_y
Sum of Squares (Variance)	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 P_{ij}$
Inverse Difference Moment	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1 + (i-j)^2} P_{ij}$
Sum Average	$\sum_{i=2}^{2Ng} i P_{x+y}(i)$
Sum Variance	$\sum_{i=2}^{2Ng} (i - \text{Sum Entropy})^2 P_{x+y}(i)$
Sum Entropy	$-\sum_{i=2}^{2Ng} P_{x+y}(i) \log[P_{x+y}(i)]$
Entropy	$-\sum_i \sum_j P_{ij} \log P_{ij}$
Difference Variance	$variance \; of \; P_{x-y}$
Difference Entropy	$-\sum_{i=2}^{N_{g-1}} P_{x-y}(i) \log[P_{x-y}(i)]$
Information measure of correlation 1	$\frac{\text{HXY} - \text{HXY1}}{\max[HX, HY]}$ where HX and HY are entropies of $P_x \; and \; P_y$
Information measure of correlation 2	$(1 - \exp[-2.0 \; (HXY2 - HXY)])^{1/2}$

Feature	Formula
Maximal Correlation Coefficient	$(second\ largest\ eigenvalue\ of\ Q)^{1/2}$ Where $Qij = \sum_k \frac{P_{ik}\,P_{jk}}{P_x(i)P_y(k)}$, $P_x\ and\ P_y\ are\ partial\ probability\ density\ functions$

Twelve Haralick features (Haralick et al., 1973) and two features defined by (Chan et al., 1997) were computed based on co-occurrence matrices constructed at four different distances in order to detect the mammographic mass (Bovis & Singh, 2000). Here suspicious regions are identified using bilateral subtraction of left and right breast image pairs. An average recognition rate achieved using ANN and 10 fold cross validation technique was 77% with overall sensitivity of 74%. The classification accuracy obtained using all the 14 Haralick features seem to be comparable with that using the first order statistical features. An effort to classify the mammographic regions into mass and non-mass categories using GLCM was done by (Khuzi, Besar, Zaki, & Ahmad, 2009). Three segmentation methods (threshold, k-mean, Otsu) were tried for extracting ROI from the region. A block processing approach was adopted for processing ROIs, so the image was divided into windows. Experiments were conducted by varying the window size (8×8, 16×16, 32×32). Then GLCM at four different directions (0º, 45º, 90º and 135º) were constructed for each window. Three GLCM features homogeneity, energy and contrast were found very useful for classification. Authors noted that using window size of 8×8 the ROC curve area obtained was 0.84 for Otsu's method, 0.82 for thresholding methods and 0.7 for k-mean clustering method. Thus by selecting the three features which were responsible for distinguishing between normal and abnormal tissues (homogeneity, energy, contrast) and approach of dividing the image into smaller portions to analyse has provided much better performance than considering all the 14 Haralick's features on the whole mammographic image.

Nithya and Santhi also used five GLCM based features namely correlation, energy, entropy, homogeneity and sum of square variance for classifying the mammographic regions into normal and cancerous categories (Nithya & Santhi, 2011). Here 200 mammograms for training and 50 for testing were used from the DDSM database. Authors report 96% accuracy, 100% sensitivity and 93% specificity values. However, no cross-validation was used for estimating the generalisation so this could merely be an effect of choosing the right ROIs for training. Another work classifying ROIs into mass and no-mass categories using GLCM features is

(Wong, He, Nguyen, & Yeh, 2012). Here four significant features were selected using the sequential forward selection technique and were correlation, angular second moment, inverse difference moment at Θ = 0 degrees and correlation at Θ = 45 degrees. Experiments on 50 ROIs from miniMIAS dataset with ANN classifier and leave one out validation provided classification accuracy of 86% which is promising as compared to (Christoyianni, Dermatas, & Kokkinakis, 1999) and (Petrosian, Chan, Helvie, Goodsitt, & Adler, 1994). Average recognition rate of 77% and sensitivity of 74% was achieved using this method.

Use of five co-occurrence matrix features namely contrast, homogeneity, inverse difference moment, entropy for classifying mammograms into three categories benign, malignant and normal can be found in (Martins, dos Santos, Silva, & Paiva, 2006). Experiments were carried out on miniMIAS database and co-occurrence matrix were computed for four different directions (0, 45, 90,135 degrees) using different grey levels (8, 6, 32, 84, 128, 256). Thus 120 measures were generated for every ROI. Then using the sequential forward selection process 8 main features were selected and Bayesian Neural Network was used for classification purpose. Mean classification rate of 86.84%, 90% sensitivity and 71.42% specificity were achieved with 180 training and 38 testing samples.

Use of 11 GLCM based features (distance = 1, 3, 5) for content based image retrieval of mammograms can be found in (Wei, Li, & Wilson, 2005). The features used here were angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy. Lesions from miniMIAS database fall in the following 6 categories: calcification, circumscribed, speculation, miscellaneous, architectural distortion and asymmetry. A maximum precision of 51% and recall of 19% were obtained when the experiments were carried out with 200 × 200 pixels ROIs from miniMIAS database with GLCM distance = 5. Christoyianni (Christoyianni et al., 1999) used statistical descriptors based on high order statistics and spatial grey-level dependence matrix (SGLDM) on miniMIAS database of mammography. They observed that the Multilayer Perceptron (MLP) classifier provided 4% better performance than the Radial-Basis-Function (RBF) classifier. The best classification accuracy achieved using Grey Level Histogram Moment features was 82.35% and using SGLDM features was 84.03%. The highest classification accuracy recorded in works using GLCM features using cross-validation technique is 86.84%. However, it is advantageous to achieve further increase in the classification accuracy. Further, the highest sensitivity and specificity value recorded with

GLCM were 90% and 71.4% respectively. The sensitivity quantifies the avoiding of false negatives and false negatives are dangerous delaying the detection and treatment for a patient. In order to have less false negatives higher values of sensitivity (ideally 100%) are expected. On the other hand specificity measures the proportion of negatives that are correctly identified. Thus higher values of specificity ensure that there are less number of false positives. False positives increase the anxiety of a patient and also increase the healthcare costs by increasing the number of biopsies. Therefore, higher value of specificity (ideally 100%) is expected in the clinical scenario. It is clear that using selective GLCM features classification performance better than first order statistical features can be obtained. However, GLCM features are computationally expensive. A computationally simple feature extraction technique which tries to improve the classification accuracy, sensitivity and specificity (Pixel N-grams) is hence proposed in this thesis.

Another statistical feature for texture representation is the Local Binary Pattern (LBP). LBP was first proposed by (Ojala & Pietikäinen, 1999). The LBP operator labels the pixels of an image by thresholding the neighbourhood of each pixel with the centre value and considering the result of this thresholding as a binary number. Similarly, the spatial relationships of all neighbouring pixels can be taken care of by computing Pixel N-grams in various directions. On the other hand proposed Pixel N-grams count the number of appearances of sequence of N consecutive pixel values in the entire image. LBP has been used for mass detection, particularly to reduce the false positive rate (Lladó, Oliver, Freixenet, Martí, & Martí, 2009). Here ROIs are divided into small squares, then LBP was used for calculating local texture descriptors of each square. The combination of these local descriptors are called spatially enhanced histograms. These global features were used along with SVM classifier for classifying the true mass from the ones being normal parenchyma. This approach was evaluated using 1792 ROIs from DDSM database. Results indicate that false positive reduction can be achieved using LBP in a much better way for lesions of large size than the smaller sized lesions. An attempt to further reduce the false positives is made by use of novel Pixel N-gram technique. An improved LBP feature extraction can be seen to be used in benign/malignant classification (J. Liu, Liu, Chen, & Tang, 2011). Here SVM classifier was used to classify 309 mammographic lesions from DDSM dataset in their study. The mass ROIs were divided into smaller windows in order to calculate the LBP features. They observed that when the window size was more than one the improved LBP operator performed better than the basic LBP operator and uniform LBP

operator. In contrast computation of Pixel N-grams do not require dividing the ROI into smaller windows which save the computation time and efforts.

LBP is sensitive to noise in the near-uniform image regions. In order to overcome this limitation a new technique was proposed by (Tan & Triggs, 2010) and is called Local Ternary patterns (LTP). Unlike LBP, it uses a threshold constant to threshold pixels into three values - 1, 0 and 1. Then the ternary pattern is split into two binary patterns to generate a descriptor double the size of LBP. Thus the feature vector size is increased, increasing the computational cost as compared to LBP. Experiments on DDSM database using SVM classifier with five-fold cross-validation concluded that the LTP outperform LBP for benign/malignant classification (Nanni, Brahnam, & Lumini, 2012). Modified LTP features have been used for benign/malignant classification by (Muramatsu, Zhang, Hara, Endo, & Fujita, 2014). The classification was carried out on 149 ROIs (91 malignant and 58 benign) from miniMIAS database. Using ANN classifier accuracy of 84.8% was achieved. The classification accuracy using LTP was found to be comparable with the selective Haralick's features.

A combined approach of using diversity index and LBP for mammographic classification into benign and malignant categories can be seen in (da Rocha, Junior, Silva, de Paiva, & Gattass, 2016). They used the DDSM database (Heath, Bowyer, Kopans, Moore, & Kegelmeyer, 2000) and the 1155 ROIs (625 malignant and 525 benign) containing the mass lesions were segmented using bounding box method. The contrast of the images was enhanced using logarithmic transformations and mean filter. Then the ROI's were represented using diversity indexes computed using histograms, second order statistics using GLCM and superior order statistics using GLRGM (Galloway, 1975) and GLGLM (Xinlily & Bent, 1994). Best results achieved using this technique were accuracy of 88.31%, sensitivity of 85%, specificity of 91.89%, positive probability ratio of 10.48, negative probability ratio of 0.16 and area under the ROC curve of 0.88. The results using combination of features are promising however, add to computational complexity and time. Further, for getting these classification results using LTP it was necessary that the lesion should be included completely in the ROI with little inclusion of background region which might be hard with lesions appearing at the edges of the breast.

Independent Component Analysis (ICA) (Hyvärinen, Karhunen, & Oja, 2004) is a method where image is represented using higher order (more than 2) statistics. In case of ICA, an image

is considered to be composed of the sum of non-Gaussian statistically independent basis images. Classification of mammographic lesions using Independent Component Analysis (ICA) can be seen in the work of (Costa, Campos, Barros, & Silva, 2007). The best accuracy was obtained using SVM classifier for DDSM database was 99.6% and for miniMIAS database was 100% discriminating benign vs malignant lesions. The results seem to be pretty good however, the authors used 50% of the images for training and 50% for training and no cross-validation was used to estimate the generalisation. This could suggest merely a chance of getting higher classification accuracy. Further, the higher order statistics incur higher computational cost.

2.2.1.2 Spectral approach

To represent texture of an image various time-frequency methods have been utilised. Here the image is decomposed into different signals by using different sampling strategies. Then the signal processing methods are used to process the images. The various signal processing techniques used for image processing are Gabor filters, Fourier transform, wavelet transform and curvelet transform.

Gabor filter banks have been used for extracting the most representative and discriminative local spatial textural features of masses at different orientations and scales especially for reducing the false positives and false negatives (Hussain, Khan, Muhammad, Berbar, & Bebis, 2012). Evaluation was conducted using 512 ROIs from DDSM database. Feature subset selection is a process of selecting a subset of relevant features so as to improve the performance of the classifier model. Feature selection was performed using (Sun, Todorovic, & Goodison, 2010) method. Then using SVM classifier and 10 fold cross-validation method the accuracy reported was 99% Experiments suggest that this method works better than other existing methods such as Local Binary Pattern (LBP) and Multiscale Spatial Weber Local Descriptor (MSWLD).

Fourier transform (Gonzalez & Wintz, 1977) is a mathematical tool which helps to represent a signal waveform by means of sine and cosine functions. Fourier transform has been used for benign/malignant classification (Ojansivu & Heikkilä, 2008). Here, local phase quantization (LPQ) approach is used which is based on the blur invariance property of the Fourier spectrum. This technique uses local phase information obtained from the two dimensional Fourier

transform. The accuracy obtained using these features was 91.6%. It has been observed that Fourier transform lacks the spatial localisation and hence perform poorly (Materka & Strzelecki, 1998). The proposed novel Pixel N-grams considers the spatial localisation and hence surmised to work better than Fourier transform based features. On the other hand Gabor filters provide better spatial localisation but there is no single resolution at which spatial structures in natural textures can be localised. Further, Gabor filters are non-orthogonal resulting in redundant features at different scales (Teuner, Pichler, & Hosticka, 1995).

To overcome the limitations of Fourier transform method, wavelet transform was proposed. Wavelet is a waveform of limited duration that has an average value of zero. Wavelet analysis is achieved by breaking up of signal into shifted and scaled versions. Classification of mammograms in normal/abnormal categories using wavelet analysis and fuzzy neural approach can be seen in (Mousa, Munib, & Moussa, 2005). In this study horizontal, vertical and diagonal coefficients from the wavelet decomposition structure are extracted. Best classification accuracy achieved with global image processing was 81.4% and with local (ROI) image processing was 77.7%. Mousa (Mousa et al., 2005) has also tried to classify the lesions into benign and malignant categories. It was observed that the best classification of masses was achieved using the features extracted from the levels 3-4. Wavelet transform has been used to extract the features from mammographic lesions in order to solve two classification problems. One is to classify in benign/malignant/normal categories and another is to classify lesions into radial, circumscribed, calcification and normal categories. Daubechies Db4 and Haar wavelets were used and experiments carried out using 100, 200, 300 and 500 coefficients. Daubechies wavelets use overlapping windows, so the high frequency coefficient spectrum reflects all high frequency changes. Haar transform decomposes a discrete signal into two sub-signals of half its length. One sub-signal is a running average or trend and the other sub-signal is a running difference or fluctuation. It is simple and fast. For the first classification normal as well as calcification lesions achieved 100% accuracy whereas, for circumscribed and radial lesions accuracy of 91.7% was reported. For the second classification experiment 100% accuracy was achieved for the benign class where as 83.3% accuracy was achieved for malignant class using 100 biggest coefficients (Ferreira & Borges, 2003). Classification into benign and malignant classes was obtained using spherical wavelet transform by (Görgel, Sertbas, & Uçan, 2015). SVM classifier was used for this classification. Generalisation was estimated using leave one out validation. The technique achieved classification accuracy of 91.4% and 90.1% for the dataset acquired from hospital of Istanbul University (Turkey) and benchmark miniMIAS

database respectively. The results demonstrate that the spherical wavelet transform performs better than the Discrete Wavelet Transform (DWT) for classification into benign and malignant classes.

Gabor filters and wavelet transforms try to capture the properties of the image parts with respect to changes in particular direction and the scale of changes. This is very useful for classification of regions with homogeneous textures. Wavelet transform is found to be superior to Gabor filters as it allows the representation of textures at suitable scale by varying the spatial resolution. Also, there is a wide range of wavelet functions, from which the suitable function according the application can be chosen. However the wavelet transform is not translation-invariant (Brady & Xie, 1996). It has been observed that the spatial frequency or spectral approach performs poorly as compared to the statistical techniques (Weszka, Dyer, & Rosenfeld, 1976).

2.2.1.3 Structural approach

In structural approaches, texture is defined by well-defined primitives and a hierarchy of spatial arrangements. Structural texture features were used for classifying the regions into cancerous or normal categories. This method was used to detect invasive lobular carcinoma. In case of invasive lobular carcinoma the tumour sizes are very small. Hence the requirement here is that the CAD should produce a very small number of false positives as the radiologist cannot verify the detection by visually inspecting the mammogram. This is achieved by finding the local minima by comparing pixel values with its 8 connected neighbouring pixels. Authors reported 50% detection rate with no false positives. However, the study was undertaken using only 48 mammograms (Lu & Bottema, 2003). Structural approach can emphasize the shape of the primitives however it is possible only for the binary images (Haralick, 1979).

2.2.2 Shape Features

Shape features are also known as geometric features or morphological features. Shape and margins of masses are two important characteristics in determining the likelihood of the masses being benign or malignant. Masses with round and oval shapes with circumscribed margins are most probably benign whereas masses with irregular shape and ill-defined, micro-lobulated or speculated margins are malignant (Kopans, 2007).

The shape feature computation can be classified broadly into two main categories: 1) Contour based methods and 2) Region based methods (M. Yang, Kpalma, & Ronsin, 2008). Contour based methods are based on the shape boundary points whereas the region based methods are based on a shape's interior points.

Another way to classify shape features is according to the processing approaches. The following are some of the commonly used approaches:

- Shape signature by 1-D function. e.g. curvature function, area function etc.
- Polynomial approximation (merging or splitting).
- Spatial interrelation – e.g. principal axis, bounding box, chain code
- Moments- e.g. Zernike moments
- Shape transform- e.g. Fourier, wavelet, R-transform

Shape signatures are computationally simple, however are sensitive to noise (D. Zhang & Lu, 2001); whereas, polygon approximation eliminates noise and leads to simplification of shapes. Polygon approximation is normally used as a pre-processing step (Selvakumar & Ray, 2013). Spatial interrelation describe the region or contour by using geometric features such as curvature, area, location, and length. This model provides compact and meaningful features. Bounding box and chain code are examples of this model. Bounding box is invariant to scaling, rotation and translation and is also robust to noise (Bauckhage & Tsotsos, 2005). A boundary or region can also be described using moments.

The concept of moment is originated from physics. Although, moment features are translation and rotation invariant, noise sensitivity and information redundancy are two major drawbacks (Celebi & Aslandogan, 2005). Though Zernike moments (Papakostas, Boutalis, Karras, & Mertzios, 2007), which are robust to noise are an exception, however at the cost of computational complexity.

Use of signal processing algorithms such as Fourier transform, wavelet transform, R-transform have also been used to represent shapes in an image. Using these transforms a shape description with different accuracy and efficiency can be achieved by choosing the number of transform coefficients. Other advantages of signal processing methods for shape description include high robustness to noise and the great coherence with human perception (Yadav, Nishchal, Gupta, & Rastogi, 2007). However, these transformations are also computationally resource intensive.

Features calculated directly from the boundary include margin speculation, margin sharpness, area, circularity measure, convexity, rectangularity, perimeter, perimeter to area ratio and acutance measure. Huo (Huo et al., 1998) used margin speculation, margin sharpness and density of mass features to classify regions into benign and malignant categories using database of 95 mammograms. Results indicate that an accuracy of 94%, sensitivity of 100% and a positive predictive value of 83% was achieved using these shape features.

Normalized Radial Length (NRL) is the Euclidean distance from the centre of the tumour to each boundary coordinates normalized by dividing by the maximum radial length. Features based on NRL are known as NRL features and include mean and standard deviation of the NRL, entropy of the NRL histogram, area ratio and zero crossing count. NRL features along with tumour circularity and patient age features were used for classification of lesions into fibroadenomas, cysts and cancer using 69 mammographic images. Generalisation was estimated using leave one out validation obtaining 82% classification accuracy (Kilday, Palmieri, & Fox, 1993). Morphological features namely perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast along with five NRL features introduced by (Kilday et al., 1993) were used for classification of regions into benign or malignant categories (Petrick, Chan, Sahiner, & Helvie, 1999). A database created by University of Michigan hospital containing 253 mammograms was used. 98% accuracy was obtained by using these features from the ROIs segmented through region growing algorithm.

Normalized chord length (NCL) is the Euclidean distance of a pair of points on the boundary of the tumour normalized by length of longest chord. The complete set of chords for a given object consists of all possible chords drawn from every boundary pixel to every other boundary pixel. Compactness is a simple measure of the contour complexity. The features based on NCL include NCL mean, NCL variance, NCL skewness, NCL kurtosis. EI-Faramawy (El-Faramawy, Rangayyan, Desautels, & Alim, 1996) used the NCL statistics for mammographic classification using 54 tumours. Classification accuracies of 95% for circumscribed/speculated, 76% for benign/malignant, and 77% for four-group classification were obtained. NCL features along with the compactness, Fourier descriptors and moment based features were used for classification using 39 ROI's from miniMIAS database (Rangayyan, El-Faramawy, Desautels, & Alim, 1997). Accuracy of 92.3% was obtained for circumscribed/speculation and 95% accuracy was obtained for benign/malignant classification.

NCL distribution possess size, translation and rotation invariance property. Further, it is stable with respect to noise or distortion in the margin boundary. However, NCL method is computationally complex and different shapes might have same chord distribution. Also, it has been observed that the NCL method is comparable with the Fourier descriptor or moment invariant methods (You & Jain, 1984).

Shape and margin features along with other features such as density, abnormality assessment rank, patient's age, subtlety value were used to classify the suspicious areas from DDSM dataset into benign and malignant categories using soft cluster neural network (Verma, McLeod, & Klevansky, 2009). The subtlety value for a lesion indicates how difficult it is to find the lesion. Its value is between 1 and 5, where 1 is "subtle" and 5 is "obvious". As opposed to hard clustering, soft clustering decides a probability distribution of an example over its classes. Then the prediction of an example is made by using the weighted average of the predictions of the classes the example is in. Soft clusters make the training process of neural network faster and avoid the iterative process. The highest classification accuracy achieved using soft cluster technique was 94%. Further, using these same features an Ensemble based technique has been applied for classification of suspicious regions into benign or malignant categories (Mc Leod & Verma, 2013). Highest classification accuracy achieved with 100 training and 100 testing masses using the ensemble technique was 98%. Ensemble increases the accuracy by reducing the variance in the prediction errors. It was observed in this study that the accuracy obtained using ensemble technique was significantly better than that using single neural network and Adaboost. Zhang (Y. Zhang, Tomuro, Furst, & Raicu, 2012) used shape features from the segmented contours with an ensemble technique for classifying lesions into benign or malignant categories. The DDSM dataset was partitioned into four subsets using patient's age. An ensemble system with four classifiers was built and trained using the different subsets. Here the mass segmentation was achieved using multiple weak segmentors. For each segmented contour, 14 shape features were computed namely: area, convex, perimeter, circularity, compactness, solidity, roughness, equivalent diameter, elongation, major axis length, minor axis length, eccentricity and extent. Using this technique classification accuracy of 72% was achieved which was better than the single classifier accuracy (56%). However, the classification accuracy (72%) was lower than other methods such as using first or second order statistical features. One of the reason could be the segmentation algorithm, as the shape features strongly depend upon the segmentation accuracy. It was observed that particularly with the old

age and small ROI size the segmentation was not very accurate resulting is poor classification performance.

Another, effort using boundary modelling and shape analysis for classification of mammograms into benign and malignant categories was done by (Rangayyan, Mudigonda, & Desautels, 2000). Here a boundary segmentation method was used to separate major portions of the boundary and to label them as convex or concave segments. An iterative procedure for polygonal modelling was used for extracting features to analyse the shape information. Thus the features of fractional concavity and speculation index along with global shape features provided classification accuracy of 82% with area under the ROC curve 79% for benign versus malignant categories. These three features resulted in 91% accuracy for circumscribed versus speculated classification. The limitation of this approach is however the inter-observer variation in drawing mass contours. Additionally, noise could lead to variation in the shape features and essentially affect the classification accuracy.

A fuzzy rule based approach was taken for classification of mammographic lesions according to the BI-RADS shape categories by (Vadivel & Surendiran, 2013). Seventeen shape and margin features (area, perimeter, max radius, min radius, Euler number, eccentricity, equivalent diameter, elongatedness, entropy, circularity1, circularity2, compactness, dispersion, thinness ratio, standard deviation of mass, edge standard deviation, shape index) were used to describe mass and used for classifying mass lesions into four categories namely: round, oval, lobular and irregular. 224 masses from DDSM database were considered and a decision tree classifier was used for the experiments. It was evident from the results that this approach was superior to existing Beamlet (Sampat, Bovik, & Markey, 2005) based approach. The classification accuracy obtained by this approach was 87.76%. Another effort in classification of mammographic lesions into round, oval, lobular and irregular masses was performed by (Mohamed, Salem, Hadhoud, & Seddik, 2016). After pre-processing, segmentation of masses was achieved by using Otsu's threshold (Otsu, 1975) and morphological operation such as erosion. Then 15 shape and margin features were computed for each mass. About 270 mammograms from the Women Health Care (WHC) program and 142 mammograms from DDSM database were used for the experiments. Three classifiers ANN, SVM and KNN were tried during the experiments out of which the ANN gave the best results. The observation was that the round and oval shapes were classified with 100% accuracy whereas, for lobular and irregular shape classification accuracy of 93% and 100% respectively for WHC database. The

classification accuracy of lobular and irregular shapes for DDSM database was reported to be 100% and 91.3% respectively.

Wei (Wei, Chen, & Liu, 2012; Wei et al., 2011) proposed a mammogram retrieval system based on similar mass lesions. Here shape and margin features of mass were extracted to represent the masses. Pre-processing step consists of brightness adjustment and median filtering in order to enhance the image and remove noise. The region growing algorithm was used for segmenting the masses. Zernike moments were used as shape features as they are reported to be good at describing shapes. Two advantages of using Zernike moments are that 1) they do not require the knowledge of precise boundary of masses and 2) contribution of each moment to image is unique and independent. It was evident that the Zernike moments outperform the Fourier, CSS and moment invariants. Margin of segmented mass was detected using Sobel operators and edge map of variation in grey levels to measure its sharpness degree was used as margin feature. Further, the density degree is represented by brightness variation of a mass = average brightness of inner region / average brightness of outer regions. This was used as another feature for distinguishing among various mass types. Results indicate that the SVM algorithm with RBF kernel function achieves best performance. (Alto, Rangayyan, & Desautels, 2005) used shape features (compactness, fractional concavity, speculation index) along with Haralick's 14 features and four edge sharpness measures to retrieve mammograms based on similar mass lesions. The retrieval accuracy of 91% and precision rate of 95% was obtained.

2.2.3 Combination of various approaches

Classification of lesions into benign and malignant categories using first and second order statistical features with MLP classifier was tested on miniMIAS database by (Uyun, Hartati, & Harjoko, 2013). It was found that the best features were 24 second order statistical features (angular second moment, contrast, correlation, variance, inverse difference moment, entropy) in four different directions (0, 45, 90, 135 degrees).

A combination of LBP, Haar wavelet features and Haralick texture features have been used for classification of mammograms into benign and malignant categories by (Joseph & Balakrishnan, 2011). By using this multi feature approach along with ANN classification accuracy of 98.6% was achieved. miniMIAS database (56 benign, 42 malignant, 42 normal)

was used for the experiments. Other approaches include template matching and model based methods. Template matching has been used for distinguishing between mass and normal tissue by (Tourassi, Vargas-Voracek, Catarious Jr, & Floyd Jr, 2003). Markov Random Field (MRF) model has been used for normal/abnormal classification (H.-D. Li et al., 1995). The results show 90% sensitivity at the expense of 1.5 false alarms per image. MRF algorithm was found to be successful in detecting small masses of less than 10mm size. Breast tumour detection can also be achieved using extreme learning approach as it has good generalization abilities and a high learning efficiency (Z. Wang, Yu, Kang, Zhao, & Qu, 2014). Wavelet modulus maxima transform, morphological operation and region growth are used for the breast tumour edge segmentation. After that, five textural features and five morphological features are extracted. Then Extreme Learning Machine (ELM) classifier was used to detect the breast tumour. It was found that the ELM based classification reflects better performance than SVM classifier as well as improved training speed. Classification of ROIs containing micro-calcifications using shape and texture features and MLP classifier can be seen in the work of (Marques, 1999). It was observed that good classification was obtained using texture features than the morphological features which was due to the segmentation errors.

It is found that shape features are superior to gradient and texture features (Haralick and wavelet) for mass classification (Mu, Nandi, & Rangayyan, 2008) as well as microcalcification classification (Soltanian-Zadeh & Rafiee-Rad, 2004). However, segmentation greatly affects the effectiveness of determining the shape and margin features of masses/calcifications and hence any errors in segmentation can diminish the classification accuracy, sensitivity and specificity values. Further, shape and boundary feature extraction is computationally complex.

2.3 Bag-of-Visual-Words Model

The aforementioned low level features namely texture and shape features represent global image features and represent the image as a whole but cannot represent the local pattern variations very well. For example, statistical features such as contrast or entropy can be same for images having objects of different size, shape and location.

The idea of capturing local pattern variations in an image gave rise to the use of Bag-of-Visual-Words models (BoVW) for image representation (Sivic & Zisserman, 2003). BoVW model was inspired by Bag-of-Words (BoW) model in the text categorization domain. This model has been proven to be efficient and is now widely deployed (Tsai, 2012; W. Yang et al., 2012).

Text documents mainly contain meaningful words and so can be represented by a feature vector of counts of various words appearing in the document. However, there are fundamental differences between text and images. Firstly, text words are discrete tokens whereas, local image descriptors are not. This necessitates techniques to generate a visual vocabulary by clustering the local feature descriptors. Vector quantisation is a common technique for this but, in contrast to the text BoW, the feature vector generated is typically high dimensional and the generation process is computationally complex (van de Sande, Gevers, & Snoek, 2011). Secondly, text is unidirectional whereas images can be read in several different directions.

A BoVW approach was first applied to video retrieval by (Sivic & Zisserman, 2003). In this approach, an image is described by a number of occurrences of different visual words. Visual words are local image patterns, which can describe relevant semantic information about an image. This model soon became popular for image retrieval and classification applications due to its accuracy (Caicedo, Cruz, & Gonzalez, 2009; Jégou, Douze, & Schmid, 2010; Nister & Stewenius, 2006; Philbin, Chum, Isard, Sivic, & Zisserman, 2007; Rahman, Antani, & Thoma, 2011; J. Wang, Li, Zhang, Xie, & Wang, 2011).

Vocabulary construction has been achieved mainly using two approaches: local patch based approach/dense sampling (Avni, Goldberger, Sharon, Konen, & Greenspan, 2010; T. Li, Mei, Kweon, & Hua, 2011) and key point based approach or sparse sampling (Csurka, Dance, Fan, Willamowski, & Bray, 2004; Pedrosa & Traina, 2013; Pelka & Friedrich, 2015). In the patch based approach, the image is divided into a number of equal sized patches by using a grid. Local features are then computed for each patch separately. Keypoints are the centres of salient patches generally located around the corners and edges. Key points are also known as interest points and can be detected using various region detectors such as the Harris-Laplace detector (corner-like structures), Hessian-affine detector (Tirilly, Claveau, & Gros, 2008), Maximally stable extremal regions or the Salient regions detector (Mikolajczyk et al., 2005). Local features are then computed for each interest point.

Some of the state-of-art local feature descriptors used for modelling texture information include Scale Invariant Feature Transform (SIFT) (Lowe, 2004), Speeded Up Robust Features (SURF) (Bay, Ess, Tuytelaars, & Van Gool, 2008), Histogram of Oriented Edges (HOG) (Dalal & Triggs, 2005), Local Ternary Pattern (LTP) (Tan & Triggs, 2010) and Discrete Cosine Transform (DCT)(Cruz-Roa, Díaz, Romero, & González, 2011). Colour hues and shape

features have also been used as local feature descriptors by some of the researchers. These local feature descriptors are briefly described below.

SIFT descriptors (Lowe, 2004), are invariant to image translation, illumination, noise, scaling, rotation and partially invariant to illumination changes. These features are robust to local geometric distortion (Mikolajczyk et al., 2005) and are the most commonly used local feature descriptors for BoVW model. However, the limitations include, high computational cost and the huge feature vector dimension (128 dimensions for each keypoint).

SURF features (Bay et al., 2008) are modified SIFT features. SURF features are high-performance, scale and rotation-invariant and they outperform SIFT features with respect to repeatability, distinctiveness, and robustness (Juan & Gwun, 2009). Computation time for calculating SURF features is reduced with the use of a fast Hessian matrix based detector and a distribution-based descriptor. SURF descriptors have been successfully applied for diabetic retinopathy lesion detection (Jelinek et al., 2013; Rocha, Carvalho, Jelinek, Goldenstein, & Wainer, 2012)

Jiang (Jiang, Zhang, Li, & Metaxas, 2015) proposed scalable method for retrieval and diagnosis of breast cancer. They used Scale Invariant Feature Transform (SIFT) features of ROIs. Contextual information of vocabulary tree is used to adjust the weights of the vocabulary tree for retrieving similar ROIs from the database. The query ROI is classified by using weighted majority votes of its best matched ROIs. Retrieval set sizes of 1-20 were considered. About 11553 ROIs from DDSM database were used for the experiments and the best classification accuracy of 88.4% was achieved by taking the retrieval set size of 5. BoVW approach also called as textons approach has been used for classification of mammograms into benign or malignant classes (Y. Li et al., 2015). Intensity and rotation normalization is performed, followed by subsampling with uniform or non-uniform intervals before classification. KNN classifier was used for classification on DDSM database (114 mass regions). A classification accuracy of 85.96% was achieved using this approach. Results indicate that the texton approach works better than other texture based methods such as wavelet, curvelet, local ternary pattern (Nanni et al., 2012), local phased quantization and Independent component analysis (Costa et al., 2007). It was observed that text on approach fails to classify the masses when mass is surrounded by glandular tissue.

BoVW approach was used for mammographic image classification using a histogram intersection method (E. Cheng et al., 2010). Histogram intersection using BoVW approach was achieved by dividing the galactographic images (23 images) into smaller parts and then generating the vocabulary of visual words with the help of vector quantization. The images can then be represented with the help of number of times the visual words in the dictionary as shown in Figure 2.2. Galactography uses mammography and an injection of contrast material to create pictures of the inside of the breast's milk ducts. Genaralisation was estimated using Leave-One-Out validation and classification was achieved using KNN and SVM classifiers. The classification accuracy of 73.3% was achieved. The results demonstrate that the histogram intersection using BoVW method outperforms the other state-of-the-art classification algorithms using l1, normalized l1 and Chi square methods. Another effort for breast density classification using BoVW can be found in (Diamant, Greenspan, & Goldberger, 2012). Classification was achieved using SVM classifier. In BoVW approach the processing steps included performing adaptive histogram equalization for enhancing the image, extracting and normalizing each patch by subtracting its mean grey level and dividing by its standard deviation, then applying principle component analysis (PCA) for dimensionality reduction, using k-means clustering for vocabulary construction and building a visual word histogram. All these steps add to the computational complexity of the system. Best classification accuracy was noted to be 85% using KNN classifier and 88% using SVM classifier.

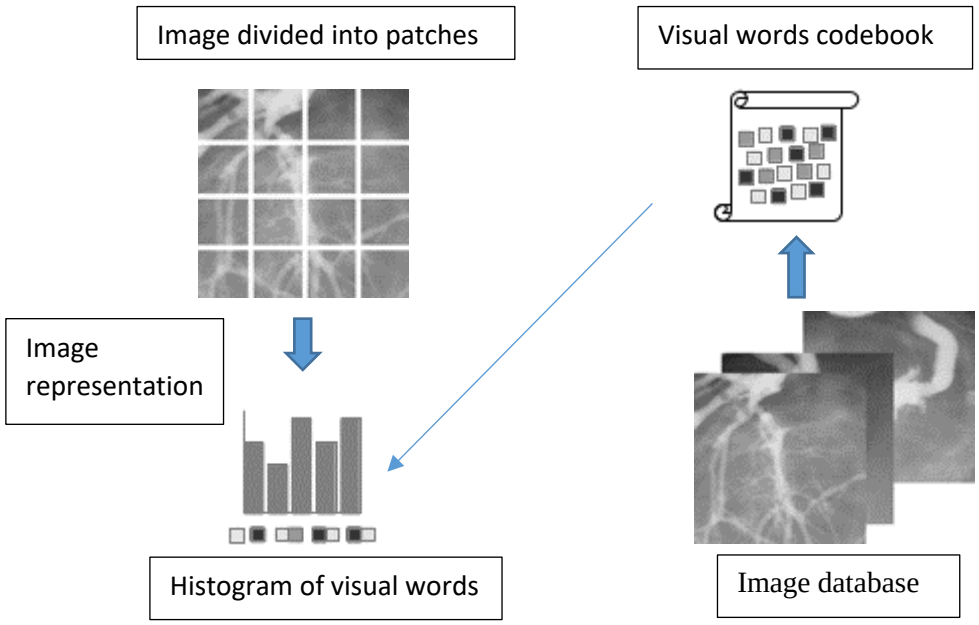


Figure 2.2 Patch based BoVW representation of galactographic images

This was achieved using 9×9 patch size and dictionary size of 100 words. The BoVW model has proven to be useful for classification (88% classification accuracy) of normal versus microcalcifications in mammographic images (Diamant et al., 2012). As opposed to the shape feature extraction, classification can be achieved using BoVW approach without segmenting the contour of each mass. Further, the local pattern variations can be modelled well with BoVW and hence superior classification performance than statistical global image features such as histogram or Haralick's features can be achieved. This rules out any segmentation errors.

Although, the BoVW model has proven to be better than models using low level global features such as histogram, Haralick's features, LBP (Tsai, 2012), it has major drawbacks. The BoVW model does not consider spatial relationships amongst visual words. Another BoVW drawback involves the high computational cost to generate vocabularies from low level features (Tirilly et al., 2008). Further, the vocabulary construction process often results in noisy words that diminish classification (Tirilly et al., 2008). The proposed Pixel N-grams method on the other hand does not require quantization step for vocabulary construction process. Hence it has some advantages over the BoVW model. One is it is computationally cost effective as the number of occurrences of sequences of grey level pixels have to be counted. The other advantage is that noisy words are not created as the quantization is not necessary. Also, the spatial relationships among the pixels are taken care of by computing Pixel N-grams in various directions. Further, the vocabulary size for Pixel N-grams is dependent upon the grey levels present in an image (usually 256) and hence is constant.

2.4 N-gram Model

Although, BoVW generates promising results in image retrieval and classification tasks; loss of spatial information and noisy words creation are two major drawbacks of this approach (Tirilly et al., 2008). The limitation of spatial information loss could be overcome by using visual N-grams (Q.-F. Zheng, Wang, & Gao, 2006). N-grams is a description obtained by grouping visual words where the arrangement between the visual words in an image is encoded. This is because the appearance of the visual words can change profoundly when they participate in relations. Further, the N-gram models for image features are simple and are able to scale up the content representation just by increasing N (Pedrosa & Traina, 2013).

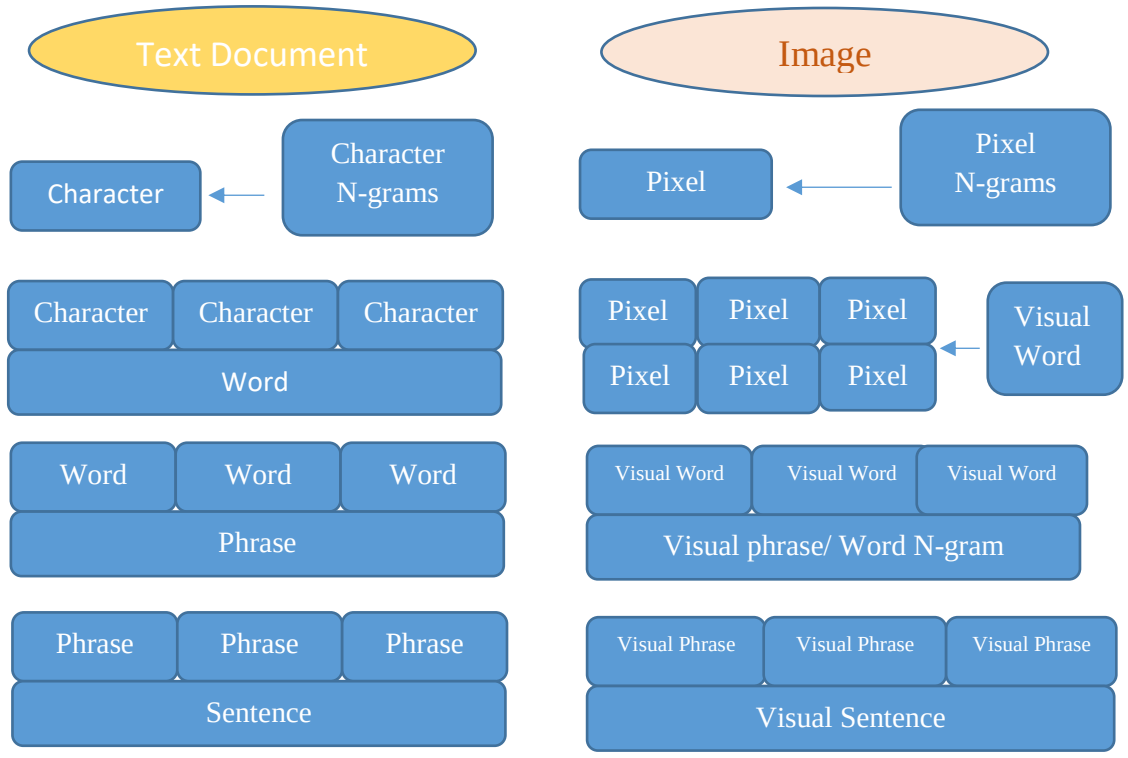


Figure 2.3 Text Vs Image N-gram analogy

Figure 2.3 shows the analogy of different types of N-gram representation for text and image. In case of text, the character N-grams can be formed by considering the N consecutive characters in a sentence. The word N-grams can be formed by considering the N consecutive words in a sentence.

Likewise, there are two approaches for visual N-grams image representation. Visual Word N-grams model consider the N consecutive visual words thus modelling the spatial relationship among visual words. In contrast, visual character or Pixel N-grams consider the N consecutive pixels in an image, modelling the relationships among the pixels. As opposed to text however, the image can be read in several different directions and hence the spatial relationships among visual words need to be considered in different directions. Further, similar to the visual sentence representation formed by many phrases in text, images can also be represented using visual sentence approach (Tirilly et al., 2008). Here, an axis is chosen for representing an image as a visual sentence, so that a) it is at an orientation fitting the orientation of the object in the image, b) it is at a direction fitting the direction of the object. The main problem is to decide the best axis for projection. The visual N-gram approaches are detailed below.

2.4.1 Visual Word N-grams

Visual Word N-grams are inspired from the word N-gram concept in text categorization or retrieval domain. In text retrieval context, word N-grams are phrases formed by a sequence of N consecutive words. E.g. 1-gram representation is composed of words such as [mass, network, reporting]. On the other hand, 2-gram is represented by sequence of two words for example [benign mass, neural network, structured reporting]. The 1-gram representation is the BoW approach. Phrases formed by N consecutive words enrich the semantics and provide more complete representation of the document thereby increasing the classification and retrieval performance. However, applying the Word N-gram concept to image is difficult as the images do not contain discrete words. For this reason local features are calculated by using one of the two main sampling strategies: keypoint based and patch based. The local features are then used to construct a vocabulary of visual words using some kind of clustering algorithms. The cluster centroids are considered visual words and are saved in the visual words dictionary. Various approaches of generating visual word N-grams are discussed below.

2.4.1.1 Dense Sampling/ Patch based N-grams

In this approach, an image is divided into small local patches using a grid. Local features are computed for each patch separately. A codebook or dictionary of visual words is then created by clustering all the patch descriptors. N-gram codebook is then developed by considering the N-consecutive visual words present in an image (Refer to Figure 2.4).

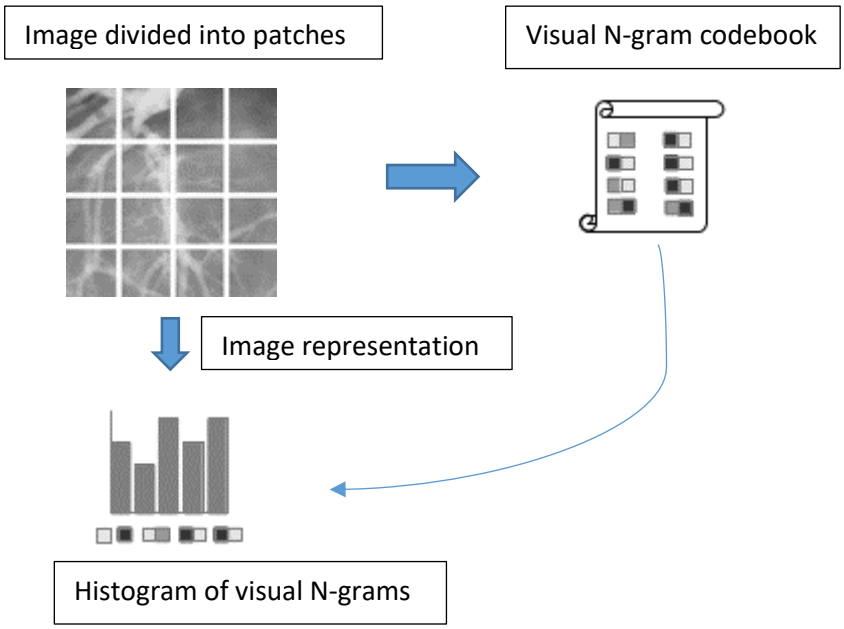


Figure 2.4 Patch based visual N-gram representation of image

The idea of N-grams using local patches was first proposed by (Zhu, Rao, & Zhang, 2002) and was called the keyblock approach. Keyblocks are similar to key words in text and using these keyblocks, images can be represented as a code matrix in which the elements are indices of keyblocks in the codebook. Uni-block, Bi-Block (horizontal, vertical, diagonal) and Tri-Block (horizontal, vertical, diagonal, triangular) configurations were used. The disadvantages of Bi and Tri-Block models are increased dimension of feature vector, large storage requirement and therefore less efficiency and retrieval performance because of highly sparse nature. However, the dimensionality of feature vector can be reduced by selecting only useful Bi and Tri-Blocks. It is reported that combination of Uni, Bi and Tri blocks result in improvement in retrieval performance. Experiments were conducted on Brodatz texture database (TDB) (Brodatz & Textures, 2009) and CDB (snapshot of images on web). Keyblock approach is compared with traditional colour histogram and colour coherent vector techniques using CDB and compared against Haar and Daubechies wavelet texture techniques using TDB. Using the keyblock approach, 12% of all relevant images were among top 100 retrieved images as compared to 9% of colour histogram and 6.5% returned by Colour Coherent Vector. Also, at each recall level keyblock approach achieved higher precision. In this study it has also been observed that the keyblock approach outperformd the Haar and Daubechies wavelet texture approaches.

Recently, local patch based N-grams were used for histopathological image classification (López-Monroy, Montes-y-Gómez, Escalante, Cruz-Roa, & González, 2013). The local

patches were represented using Discrete Cosine Transform (DCT) features. Here, the main idea was to produce N-grams ignoring the orientation in which they appear. Visual N-grams that have the same order but different orientation (e.g., if an image is rotated), like 12-65-654 and 654-65-12 are considered same, thus making the N-gram features rotation invariant. Another main idea in this study was to combine the N-gram features such as 1+2gram, 1+2+3 gram and 1+2+3+4 gram. The 1+2 gram produced the highest classification accuracy of 64.31%. The reason is because longer sequences produce large vocabulary resulting features specific to the particular image and thus make it hard for the classifier to generalize. Results re-enforce the fact that use of N-grams outperform the BoVW technique. Composing simple image descriptions using the patch based N-grams can be seen in (S. Li, Kulkarni, Berg, Berg, & Choi, 2011). It is observed that key point based samplers such as Harris-Laplace work well for small numbers of sampled patches; however, they cannot compete with uniform random patch based sampling using larger numbers of patches for best classification results (Jurie & Triggs, 2005).

2.4.1.2 Sparse Sampling/ Keypoint based N-grams

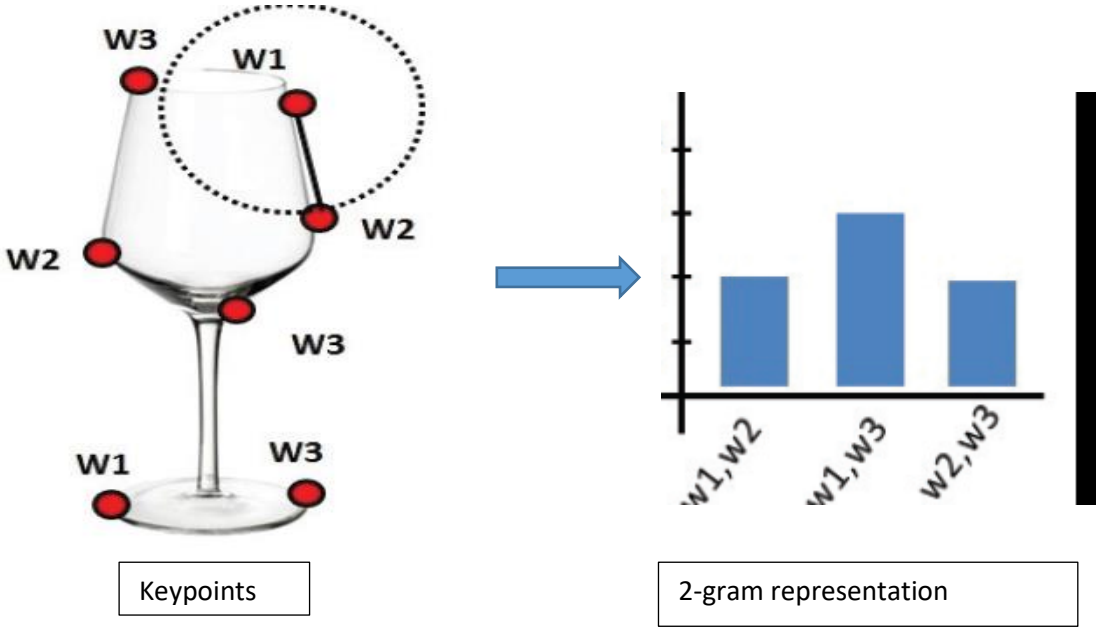


Figure 2.5 Keypoint based N-gram representation of image

Keypoints/Interest points are the points of local maxima and minima of difference of Gaussian function (Lowe, 2004). These keypoints are described with the help of SIFT features and clustered for construction of visual vocabulary. The centroids of the clusters represent visual words. The N-gram dictionary is then created considering N neighbouring visual words in all possible directions (Pedrosa & Traina, 2013). Figure 2.5 shows keypoint based 2-gram

representation of the image. In this figure w1, w2, w3 are the keypoints. The spatial relationship between these keypoints in all directions is considered while computing the number of occurrences of the 2-grams for example w1w2, w2w3 and w3w1.

It is evident that as N increases, a more complete representation of an image (spatial relationship with every pixel to every other pixel is modelled) is generated. Here, authors have used 1, 2 and 3 grams to analyse retrieval precision as well as classification accuracy on various databases namely Corel 1000 (J. Li & Wang, 2003), Lung database, Medical Image Exams database, Texture database. These experiments show that the visual word N-grams (bag-of-visual-phrases) approach improved retrieval precision up to 44% and classification accuracy up to 33% compared to BoVW approach. However, the use of visual words to represent an image in this way may involve a loss of fidelity to visual content since two local features associated with the same visual word are used in the same way to construct the image signature, whether they are identical or noticeably different. An approach for generating a more realistic image signature considering the differences between textual words and visual words can be seen in the work of (Bouachir, Kardouchi, & Belacel, 2009). Some more examples of the use of keypoint based N-grams are large scale image retrieval (Dai, Sun, Wu, & Yu, 2013), automatic learning of visual phrases (S. Wang et al., 2013), classification of images in Caltech dataset (Mukanova, Hu, & Gao, 2014) and biomedical image classification (López-Monroy et al., 2013; Pedrosa, Rahman, et al., 2014; Pedrosa, Traina, & Traina, 2014).

For visual characterisation, the frequency of occurrence of visual words as well as the spatial information between the visual words is equally important. A major challenge in using word N-grams is the dimensionality and hence the computational cost. It is clear that the number of all possible combinations of N-grams increases exponentially with N. That is, given a dictionary with m words, the number of all possible N-grams is m^N.

A novel effective and efficient technique to extract the frequency and appearance of visual words has been proposed in (Pedrosa, Rahman, et al., 2014). In this approach 2-grams were generated by placing a circular region over each keypoint. All pairs of words in this region formed with the centre point are 2-grams. Two bags of 2-grams are then generated. One bag for 2-grams with angle within [-135, 135] and [-45, 45] and another bag for 2-grams with angle within the interval [135, 45] and [-135,-45]. Then the frequency of 2-grams for each bag according to dictionary of 2-grams is noted. This is called as bag-of-2-grams approach. The

results demonstrate that the classification accuracy is improved by 6.03% as compared to the BoVW approach. Further, this approach computes the Shannon entropy over a random "bunch" of 2-grams and demonstrates that the dimensionality can be significantly reduced.

2.4.1.3 Colour N-grams

Colour features have been used for Content Based Image Retrieval (CBIR) because they can be easily extracted and are powerful descriptors for images (Y. Liu, Zhang, Lu, & Ma, 2007). Colour histograms representing relative frequency of colour pixels across the image are common for CBIR. However, they only convey global image properties and do not represent local colour information. In the Colour N-grams approach, an image has been represented with respect to a codebook, which describes every possible combination of a fixed number of coarsely quantised colour hues (R. Rickman & Rosin, 1996). This allows comparison of images based on shared adjacent colour objects or boundaries. N-gram samples were taken to be 25% of the total number of pixels in an image. The dataset included 100 general colour images of faces, flowers, animals, cars and aeroplanes. The results were compared with the approach adopted by (Faloutsos et al., 1994). The average rank of all relevant images was reported to be 2.4 as compared to the 2.5 of the baseline. Also, the number of relevant images missed was 1.9 as compared to 2.1 of the baseline. The limitation of this study is that the quantisation of the hues does not match the sensitivity of the human colour perception model. Another limitation was the very small database used. However, further work has demonstrated that this approach could also be used for very large databases (R. M. Rickman & Stonham, 1996). Moreover, this approach is less sensitive to small spectral differences and is not prone to colour constancy problems.

2.4.1.4 Shape N-grams

The concept of N-gram has been used to group perceptual shape features to discover higher level semantic representation of an image (Mukanova et al., 2014). Here, low-level shape features were extracted and perceptually grouped using the Order Preserving Arctangent Bin (OPABS) algorithm advanced by (Hu and Gao, 2010). This is based on Perceptual Curve Partitioning and Grouping (PCPG) model (Gao & Wong, 1993). In this PCPG model, each curve is made up of Generic Edge Tokens (GET) connected at Curve Partitioning Points (CPP). Each GET is characterized by monotonic characteristics of its Tangent Function (TF) set. The extracted perceptual shape descriptors are categorized as one of eight generic edge segments.

Gao and Wang's model is based on Gestalt theory of perceptual organisation which states that humans perceive the objects as a whole. The authors define shape N-gram as a continuous subsequence of GETs connected at CPP points. There were three main cases of how the GETs were connected at CPP. The first references a curve segments connected to another curve segment (CS-CS); the second is a line segment connected to line segment (LS-LS), and third is curve segment connected to line segment (CS-LS). Here, four N-gram based perceptual feature vector were proposed, which encode local and global shape information in an image. The Caltech256 dataset was used for classification experiments (Griffin, Holub, & Perona, 2007). Results show that the combination of shape N-grams with conventional SIFT vocabulary achieve around 8% higher classification accuracy as compared to SIFT based vocabulary alone.

Further, the development of CANDID (Comparison Algorithm for Navigating Digital Image Database)(Kelly & Cannon, 1994) was inspired by the N-gram approach to document fingerprinting. Here a global signature was derived from various image features such as localised texture, shape or colour information. A distance between probability density functions of feature vectors was used to compare the image signatures. Global feature vectors represent single measurement over the entire image (e.g dominant colour, texture). Whereas, the N-gram approach allows for the retention of information about the relative occurrences of local features such as colour, grey scale intensity or shape. Use of probability density functions can reduce the problem of high dimensions, however they are computationally more expensive than histogram based features (Kelly, Cannon, & Hush, 1995). It was observed that subtracting a dominant background from every signature prior to comparison does not have any effect while using true distance function; whereas, considering a similarity measure such as nSim(I1,I2), dominant background subtraction has a dramatic effect. Experimental results show good retrieval precision.

2.4.2 Visual sentence approach

A new representation of images that goes further in the analogy with textual data, called visual sentences, has been proposed by (Tirilly et al., 2008). A visual sentence that allows visual words to be read in a certain order. An axis is chosen for representing an image as a visual sentence, so that a) it is at an orientation fitting the orientation of the object in the image, b) it is at a direction fitting the direction of the object. The keypoints were then projected onto this

axis using orthogonal projection. In this work SIFT descriptors were used and keypoints detection was achieved using Hessian-affine detector. The main problem here was to decide the best axis for projection. Experiments include five different axis configurations: One PCA axis, two orthogonal Principal Components Analysis (PCA) axis, ten axis obtained by successive rotation of 10 degrees of main PCA axis, X axis and finally one random axis. Results show that the approach with X-axis outperforms those with the PCA axis on classification tasks. This is because the PCA axis is biased by background clutter. However, PCA axis takes spatial relations into account and outperforms the random axis or the multiple axis configurations. They also observed that the keypoint detection algorithm is not perfect and because images have different backgrounds, the vocabulary construction process results in many synonymic words. Therefore, N-grams longer than length of four produce bad results.

2.4.3 Contextual bag-of-words

Two relations between local patches in images or video key frames can be important for categorization. First, there is the semantic conceptual relation between patches. That is relation of appearing on the "same part", "same object" or "same category". For example "wheel of a motorbike", "window of a house", "eye of human". Further, semantic relations can be interpreted in multiple levels; for example patches of same scene, object and object parts and so on. Second is the spatial neighbourhood relation. Patches when combined together to form a meaningful object or object part are considered as having spatial neighbourhood relation. These two types of relations were called as contextual relations. Traditional BoW model neglects the contextual relations between local patches. Nevertheless, it is well known that the contextual relations play an important role in recognising visual categories from their local appearance. On the 15 scene database, the classification accuracy using contextual-bag-of-words was found to be significantly better than the traditional BoVW model (T. Li et al., 2011).

Thus in the N-grams model, image is represented with the number of occurrences of the visual phrases present in it. This is a more powerful representation than BoVW model as it takes into account the spatial relationship between consecutive visual words. However, couple of limitations still exist for this approach. Noisy words and hence noisy phrases can be created due to the quantisation step in the vocabulary construction process. Complexity is increased as compared to the BoVW model resulting in higher computation times. Further, word N-grams considerably increase the dimensionality of the feature vector.

2.5 Convolutional/Deep Learning Neural Networks

Traditionally, problem dependent features were used to represent the content of images and used for classification or retrieval purposes. The most discriminating features can then be selected using the filter approach, wrapper approach or sequential forward selection approach. An alternative approach of using neural networks for automatically finding the effective features has been proposed recently and is seen to be quite successful.

Convolutional neural network (CNN) is a neural network which shares connections between hidden units leading to low computational time and translational invariance properties. Use of convolutional neural networks for learning the features for mammographic mass lesions is found in (Arevalo, González, Ramos-Pollán, Oliveira, & Lopez, 2015). Breast cancer digital repository (Jalalian et al., 2013) database was used here and the results show area under ROC curve of 86% which seems to outperform other state-of-the-art techniques such as Histogram of oriented gradients (HOG) and Histogram of gradient divergence (HGD) (79.9%). A cascade of two level deep convolutional neural networks and random forest classifiers were used for mammographic mass detection (Dhungel, Carneiro, & Bradley, 2015). The method was tested on DDSM and INbreast (Moreira et al., 2012) database of mammography. A true positive rate of 0,96 at 1.2 false positives per image on INbreast and true positive rate of 0.75 at 4.8 false positives per image on DDSM database was achieved using this cascade approach. Further, use of CNN for estimating patients risk of developing breast cancer can be seen in the work of (Carneiro, Nascimento, & Bradley, 2015). Experiments on DDSM and Inbreast database conclude that the area under the ROC curve (benign/malignant) over 0.9 and volume under ROC surface of 0.9. They also show that the CNN pre-trained using the computer vision database (Imagenet) can be used in medical image applications. Deep belief networks have recently been used for breast cancer detection (Abdel-Zaher & Eldeib, 2016). The technique was tested on Wisconsin breast cancer dataset (WBCD). Classification accuracy of 99.68% sensitivity of 100% and specificity of 99.47% was obtained using this approach. The limitation of the deep learning neural networks is that it requires huge amount of data for training.

2.6 Novel Visual Character N-grams/ Pixel N-grams

To overcome the aforementioned drawbacks of BoVW and Word N-grams we propose a novel computationally less expensive visual character N-gram/Pixel N-gram model for classification of mammographic lesions. The model is inspired from the character N-gram model in text categorization domain.

Essentially, character N-grams are formed by sequence of N consecutive characters. For example, the 3-grams in the phrase "his pool" "his, is_, s_p, _po, poo, ool" and the 4-grams are "his_, is_p, s_po, _poo, pool". Character N-grams have shown to be able to capture information at various levels: lexical, word-class and use of punctuation marks. Furthermore, the character N-grams are language independent, robust to grammatical errors and do not require any text pre-processing (tokenizer, lemmatizer or other NLP tools) (Millar et al., 2006; Kanaris et al., 2007). In languages such as Chinese, where there are no specific word boundaries character N-grams have been proven to be very useful for document retrieval.

Along the same lines it is difficult to find specific word boundaries for an image. Therefore, an image can be represented using sequences of adjacent visual characters/pixels. This is called *Pixel N-gram/Visual Character N-gram Model* for representation of image. In this model every grey level value of a pixel is considered a visual character. However, application of character N-gram model for images is not straightforward. While text documents have a single spatial direction, images are two dimensional and the sequence of feature descriptors can be obtained in different orientations (vertical, horizontal or at an angle of θ degrees). The images are then represented with the histogram of sequence of intensity levels of pixels. By looking at the N adjacent grey levels with the help of sliding window the number of occurrences of sequence of N grey level values can be counted.

As compared to the statistical or spectral features the Pixel N-gram features only involves counting the appearances of the sequences and hence are computationally cost effective. Further, in Haralick's features based on co-occurrence matrix the spatial relationship between two adjacent pixels are modelled; whereas, Pixel N-grams can model the spatial relationships between N adjacent pixels improving the image representation and hence the classification performance.

Further, shape features are quite useful for classification of images however, the shape representation requires segmenting the region of interest/object under consideration. Also, the classification accuracy using shape features is highly dependent upon the accuracy of the segmentation algorithms. In case of dense breast tissue, segmenting the region of interest is challenging. The advantage of Pixel N-grams over shape features is that they do not require segmentation of ROIs. Moreover, shape feature computation is computationally complex whereas, the Pixel N-grams are easy to compute requiring less computational overhead.

The spatial relationships among the pixels can be taken care of by computing the Pixel N-grams in different directions thus overcoming the spatial relationship limitation of the BoVW approach. Further, the Pixel N-grams do not require the quantisation step for vocabulary construction and hence noisy words/phrases problem from BoVW or word N-grams can be eliminated. Also, in the BoVW or word N-grams approach some information is lost due to the vocabulary construction process which can affect the classification performance. On the other hand Pixel N-gram approach considers every pixel value in an image and hence no information is lost. Further, the Pixel N-grams are easier to compute than the SIFT features used for the BoVW or visual words approach, thereby requiring less computational cost.

Also, it is evident that character N-grams minimise the problem of sparse data to a great extent (major problem of dimensionality while using word N-grams) (Kanaris et al., 2007). Obviously, there are much fewer character combinations than word combinations, and hence less N-grams have zero frequency. To the best of our knowledge, the concept of character N-grams has not been explored for image classification applications.

Detection of mammographic masses is a challenging problem due to their large variation in texture, shape, boundary and their low signal to noise ratio compared to the surrounding breast tissue**.** It is surmised that the Pixel N-gram counts would be able to model the shape and texture properties of an abnormality in a mammographic image.

In order to check if the Pixel N-grams can distinguish between various textures, Pixel N-grams was applied to texture image classification (P. Kulkarni, Stranieri, A., Ugon, J., Aug 2016). Further, Pixel N-grams approach was also tested on shapes classification (P. Kulkarni, Stranieri, & Ugon, 2016). Further, the Pixel N-gram approach was used for classification of secondary digital mammograms (mammograms digitised with the help of a scanner: miniMIAS

database) (P. Kulkarni, Stranieri, Kulkarni, Ugon, & Mittal, Feb 2014, Mar 2014) and primary digital mammograms (LakeImaging dataset). Classification of ROIs into normal/abnormal categories was performed and then the fine-grained classification of mass lesions into circumscribed/speculation and normal categories was performed.

2.7 Chapter Summary

Imbalanced learning is the process of learning from data sets where the number of samples belonging to one class is significantly lower than those belonging to the other classes. This has a major impact on machine learning as the algorithms do not take into account the class distribution / proportion or balance of classes. The summary of all works described in this chapter is given in Table2.2.

Table 2.2 Summary of all related works

Author	Year	Dataset	Features	Classification	Classifier	Accuracy (%)	Balanced data?
Brake et al.	2000	DDSM (772 images)	Contrast, intensity and location	Circumscribed masses and stellate lesions	-	75	-
Bovis & Singh	2000	miniMIAS	Haralick's features	Mass and no mass	ANN	77	N
Khuzi et al.	2009	miniMIAS	GLCM	Mass and no-mass	Otsu method	84	N
Lladó et al.	2009	DDSM	Local Binary Pattern	Mass and no mass	SVM	90.6	N
Hussain et al	2012	DDSM	Gabor filter	Mass and no mass	SVM	99	Y
Nithya and Santhi	2011	DDSM	correlation, energy, entropy, homogeneity and sum of square variance	Normal Abnormal	Neural Network	96	Y
Wong et al.	2012	miniMIAS	GLCM (4 selected features)	Mass and Normal	ANN	86	Y
Martins et al.	2006	miniMIAS	GLCM (8 selected features)	Benign, Malignant and Normal	Bayesian Neural Network	86.84	N
Christoyiann i et al.	1999	miniMIAS	Grey Level Histogram Moment	Benign malignant	MLP	82.35	Y
Liu et al.	2011	DDSM	LBP	Benign Malignant	SVM	66.14	Y
Nanni et al.	2012	DDSM	Local Ternary Pattern	Benign Malignant	SVM	97.0	-

Author	Year	Dataset	Features	Classification	Classifier	Accuracy (%)	Balanced data?
Muramatsu et al	2014	miniMIAS	Modified Local Ternary Pattern	Benign Malignant	ANN	84.8	N
da Rocha et al.	2016	DDSM	Diversity index and LBP	Benign Malignant	SVM	88.31	N
Costa et al.	2007	miniMIAS	Independent component analysis	Benign Malignant	SVM	97	-
Mousa et al.	2005	miniMIAS	Wavelet Transform	Benign Malignant	Fuzzy Neural	81.4	N
Ferreira & Borges	2003	miniMIAS	Haar Wavelet Transform	Benign Malignant	Nearest Neighbour	83.3	-
Görgel et al.	2015	miniMIAS	Spherical Wavelet Transform	Benign Malignant	SVM	90.1	Y
Huo et al.	1998	95 images	margin speculation, margin sharpness and density	Benign Malignant		94	-
Petrick et al.	1999	University of Michigan hospital	perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast, NRL	Benign Malignant	ANN	98	-
(El-Faramawy et al.	1996	54 tumours	NCL mean, NCL variance, NCL skewness, NCL kurtosis	Benign Malignant		76	-
Rangayyan	1997	miniMIAS	Fourier descriptors and moment based features	Benign Malignant	Jackknife method Mahalanobis distance	95	N
Verma et al	2009	DDSM	density, abnormality assessment rank, patient's age, subtlety	Benign Malignant	Soft Cluster NN	94	Y
Joseph & Balakrishnan	2011	miniMIAS	LBP, Haar wavelet features and Haralick	Benign Malignant	ANN	98.6	-
Zhang et al	2012	DDSM	14 shape features	Benign Malignant	Ensemble	72	Y
Mc Leod et al	2013	DDSM	density, abnormality	Benign Malignant	Ensemble	98	-

Author	Year	Dataset	Features	Classification	Classifier	Accuracy (%)	Balanced data?
			assessment rank, patient's age, subtlety				
Jiang et al.	2015	DDSM	SIFT	Benign Malignant		84.4	-
Li et al.	2015	DDSM	BoVW	Benign Malignant	KNN	85.96	Y
Rangayyan	1997	miniMIAS	Fourier descriptors and moment based features	Circumscribed / Speculation	Jackknife method Mahalano bis distance	92.3	N
Rangayyan et al	2000	miniMIAS	Boundary modelling	Circumscribed /Speculation	Jackknife method Mahalano bis distance	91	Y
Vadivel & Surendiran	2013	DDSM (224)	17 shape and margin features	round, oval, lobular and irregular	Decision Tree classifier	87.76	N
Mohamed et al.	2016	DDSM (142)	15 shape and margin features	round, oval, lobular and irregular masses	ANN	91.3	N
Cheng et al.	2010	Galactographic	Histogram Intersection	Fatty, glandular	SVM	73.3	Y
Diamant et al.	2012	Mammograms	BoVW	normal micro-calcifications	SVM	88	Y

In this chapter, need for the automated classification of mammograms is explained. The application of classification of mammograms for various purposes such as cancer detection, cancer diagnosis, training of radiologists and medical research has been elaborated. The most important thing in the classification of images is the features that represent the image. Various features used for classification of mammographic images have been outlined. The features include the global low level features such as texture and shape features. The texture features are mainly grouped under 3 main approaches, namely statistical (first order, second order and higher order statistics), spectral (Fourier transform, wavelet transform, curvelet transform) and structural (texton approach). Shape features are mainly classified as contour based features or region based features. The statistical texture features and shape features are both computationally expensive. Also, they are global image features and hence are not able to model the local pattern variations very well. The BoVW model inspired from the BoW model in text retrieval is explained later which is shown to outperform the global texture and shape features. Few of the limitations of BoVW model include computational complexity, noisy

words creation and ignoring the spatial relationship between visual words. The spatial relationships are taken into consideration in the visual N-grams model inspired from word N-grams in text categorization. Different approaches for computing the visual word N-grams such as patch based N-grams, keypoint based N-grams, colour N-grams andshape N-grams are discussed. Higher up the hierarchy, visual sentence based image representation is outlined along with its limitations. Contextual bag-of-visual-words which considers contextual relationship as well as spatial relationships between visual words is discussed. Further, deep learning neural networks (state-of-the-art) approach for image classification is explained briefly. Finally, the proposed Pixel N-gram approach is detailed with its advantages over other existing techniques. Various experiments are carried out in order to evaluate the efficacy of Pixel N-grams for image classification applications. Texture and shape are two important characteristics of the lesions in mammograms. The experiments are conducted on texture images, shape images and then on the mammographic images. The methodology for the experiments will be detailed in the Chapter 3. The experimental results and analysis on the texture and shape dataset are detailed in Chapter 4, whereas, the experimental results and analysis on miniMIAS and LakeImaging database of mammography are detailed in Chapter 5.

3 Research Methodology

It is clear from Chapter 2 that there is a need for computationally efficient algorithm for classification of mammographic lesions that can maintain high classification accuracy, high sensitivity as well as high specificity (almost no misses). In this research, a novel Pixel N-gram model is proposed for classification of mammograms.

This chapter presents methodology used for fulfilling the aims of this research project. An empirical approach has been adopted and a series of classification experiments have been designed to test the efficacy of Pixel N-grams for mammographic lesion classification. Experiments have also been designed to compare the performance with the state-of-the-art feature extraction techniques. A step by step procedure with all the parameters and software programs used for the experiments is described in this chapter. The datasets and the evaluation metrics used for comparing the performances have also been explained in detail.

The main aim of this research is to classify mammographic lesions using the Pixel N-grams technique. Two types of classifications were observed. First was the normal/abnormal classification which is helpful for automated screening and the second is fine-grained classification (circumscribed lesion, speculated lesion and normal) which is useful for automated diagnosis of lesions. Two mammographic datasets were used for evaluating classification performance. These two databased are, a benchmark dataset miniMIAS (Suckling et al., 1994) and a dataset specially prepared for this project (provided by LakeImaging Pvt. Ltd. Inc.). The Lakeimaging dataset was provided by Victoria based radiology firm Lake Imaging and consisted of high resolution truly digital mammograms[6] generated via patient scans during 2013-2014. The miniMIAS dataset consists of secondary digital mammograms[7].

As per BIRADS (Breast Imaging Reporting And Data System) standard mammographic lesions are characterised by texture and shape of the lesions (D'Orsi, 2013). Texture is defined as a quantitative measure of arrangement of intensities in an image and can be modelled using

[6] Truly digital mammograms/ primary digital mammograms are digital mammograms directly generated with the help of advanced imaging equipment.
[7] Secondary digital mammograms are conventional film based mammograms digitised with the help of a scanner.

various approaches namely statistical, spectral and structural (J. Zhang & Tan, 2002). Texture classification experiments were designed with the intention of analysing to what extent Pixel N-grams can distinguish between various textures. Further, lesions in mammographic images are of different shapes and sizes. They could be located at different locations in an image. Also, mammograms produced at different imaging stations could be of different resolutions. The effectiveness of the Pixel N-grams approach for shape classification has been analysed with the help of a specially prepared dataset of binary images of three basic shapes (circle, triangle and square). Experiments were conducted to find out the extent to which classification using Pixel N-grams is independent of size and location of a shape in an image. Further, classification of basic shapes with different image resolutions using Pixel N-gram features was also assessed. A model that classifies mammograms containing lesions of various shapes and sizes regardless of image resolution, from those that do not contain lesions will potentially lead to classification systems that are clinically useful in practice.

All the classification experiments were conducted using Weka 3.6 data mining software developed at University of Waikato (Weka, 2011). The machine used for all experiments was i5-4210U CPU @2.90GHz PC with windows 10 (64 bit) operating system. All the algorithms for grey scale reduction, Pixel N-gram feature extraction, intensity histogram feature extraction, co-occurrence matrix based feature extraction as well as computational time comparison and shape database creation were implemented using Matlab 7.9 Software.

3.1 Research Approach

An experimental/empirical methodology has been adopted in this study in order to analyse the effectiveness of Pixel N-gram features for image classification. Essentially, classification on four main types of images was studied; texture images, shape images, primary mammographic images and secondary mammographic images. The classification performance of the Pixel N-gram features was analysed using a quantitative approach in order to compare Pixel N-gram approach with existing state-of-the-art feature extraction techniques used for mammographic classification.

Statistical measures were used to evaluate the performance of the classification algorithms. These include classification accuracy, sensitivity, specificity, precision, false positive rate and

Fscore (Sokolova & Lapalme, 2009). For example, consider the confusion matrix for abnormal/normal classification (See Table 3.1).

Table 3.1 Confusion matrix for normal/abnormal classification

		Actual Output	
		Abnormal	Normal
Predicted Output	Abnormal	True Positive (TP)	False Positive (FP)
	Normal	False Negative (FN)	True Negative (TN)

Classification accuracy is the number of correct predictions out of total predictions made. It is given by the equation 3.1, where t = number of correctly predicted/classified instances = TP + TN and N = total number of instances to be classified = TP +TN + FP + FN.

$$Accuracy = \frac{t}{N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Other than the Accuracy, Sensitivity and Specificity are the two very important factors to consider when it comes to detection of certain disease from clinical perspective. The higher numerical value of sensitivity suggests lower number of false-positive results. Sensitivity is also known as true positive rate or recall and is given by equation 3.2 A test with high sensitivity tends to capture all possible positive conditions without missing anyone.

$$Sensitivity/Recall = \frac{TP}{TP + FN} \tag{3.2}$$

Specificity measures the proportion of negatives that are correctly classified. Specificity is also known as true negative rate and is given by equation 3.3. The higher value of specificity indicates a lower number of false negatives. Thus, high specificity is better for ruling out a particular disease condition. High values of both specificity and sensitivity are expected for diagnostic purposes.

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

Precision is another important evaluation metric useful for classification and retrieval applications. Precision tells how many of the positively classified instances were relevant and is given by equation 3.4.

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

The False Positive rate is the ratio of number of instances misclassified as positive to the total number of actual negative instances and is given by the equation 3.5.

$$FP\ Rate = \frac{FP}{TN + FP} \tag{3.5}$$

Fscore is a harmonic mean of precision (positive predictive value) and recall (true positive rate or sensitivity). It is calculated using the equation 3.6.

$$F_{score} = 2 \times \frac{Precision\ \times Recall}{Precision + Recall} \tag{3.6}$$

3.2 Datasets Used

Performance of Pixel N-gram features was evaluated using four different datasets. These datasets include UIUC texture dataset, basic shapes dataset, miniMIAS mammographic dataset and LakeImaging mammographic dataset.

Table 3.2 Aspects of Datasets used for this study

Dataset Name	Number of examples	No. of Predictive attributes	No. of Classes	% examples in majority class
UIUC Texture	1000	318	25	33.3%
Basic Shapes	240	07	3	33.3%
miniMIAS	270	147	3	76.7%
LakeImaging	80	150	3	50.0%

3.2.1 UIUC texture dataset

To determine the effectiveness of Pixel N-grams for the classification of texture images, a benchmark texture database available publically at http://www-cvr.ai.uiuc.edu/ponce_grp (Lazebnik, Schmid, & Ponce, 2005) has been used.

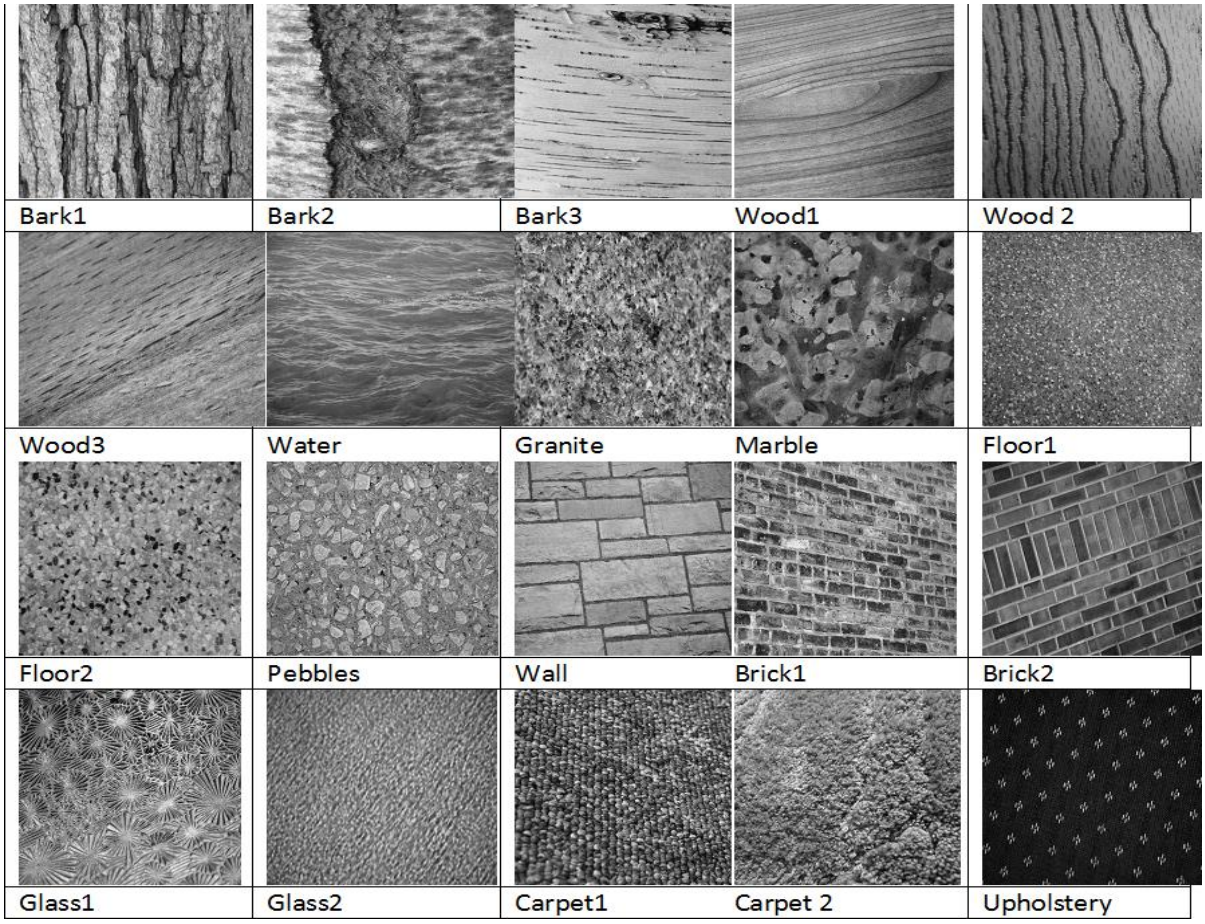


Figure 3.1 Sample images from UIUC dataset

The database consists of 1000 texture images of 25 different classes. Each class has about 40 images. All images are in Grey scale JPG format with resolution of 640 x 480 pixels. Figure 3.1 shows some of the sample images from UIUC dataset. The database includes surfaces whose texture is due mainly to reflectivity variations (e.g.,wood and marble), 3D shape (e.g., gravel and fur), as well as a mixture of both (e.g., carpet and brick). Significant viewpoint changes and scale differences are present within each class and illumination conditions are uncontrolled. Additional sources of variability exist such as non-planarity of the textured surface (bark), significant non-rigid deformations between different samples of the same class (fur, fabric, and water), inhomogeneity of the texture patterns (bark, wood, and marble), and viewpoint-dependent appearance variations (glass).

3.2.2 Basic shapes dataset

Mammographic lesions are of different shapes, sizes and they can be located at different locations in a mammogram. In order to analyse how well classification using Pixel N-grams work for the shapes a dataset of basic shapes was prepared.

This basic shapes dataset consist of binary images of three basic shapes (80 circles, 80 triangles, 80 squares). Three shapes square, triangle and circle were selected because these are geometrically diverse and are basic. These shapes are of different sizes. The shapes are located at different locations. All these images are of 512 × 512 pixels resolution. These images were used for testing the size and location invariance of Pixel N-gram features for classification of shapes.

The shapes were constructed as solid white on black background as the breast lesions appear with high intensity than the surrounding tissue on mammographic images. Sample images from this dataset are shown in the Figure 3.2

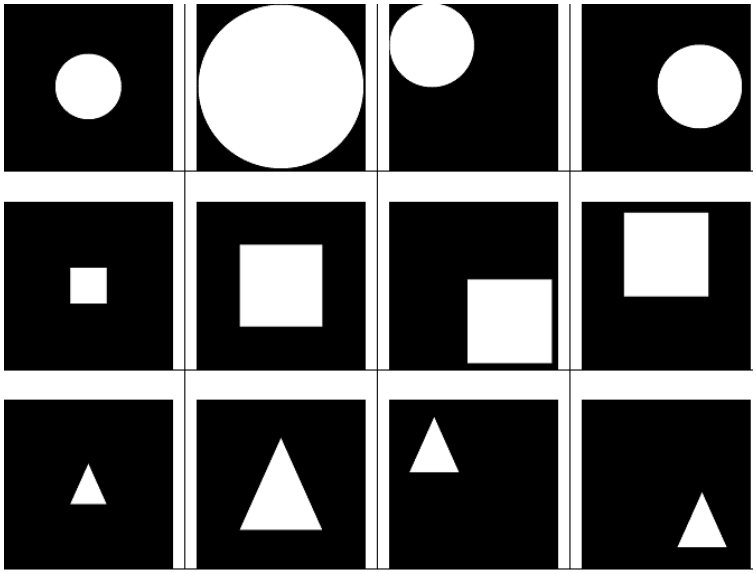


Figure 3.2 Sample images from basic shapes dataset

Further, images of 10 different resolutions (for every shape image) were generated to test the effect of varying resolution on classification performance. The resolutions used were 512 × 512, 1024 × 1024, 1536 × 1536, 2048 × 2048, 2560 × 2560, 3072 × 3072, 3584 × 3584, 4096 × 4096, 4608 × 4608, 5120 × 5120 pixels.

3.2.3 miniMIAS dataset of mammography

The Mammographic Image Analysis Society, UK has provided Mini-MIAS as benchmark medical image database for research purposes (Suckling et al., 1994). The mammograms have been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024 × 1024 pixels. The images are in portable grey map (PGM) format which can be directly read using Matlab software. All the images have grey levels ranging from 1 to 256. The distribution of grey levels in the miniMIAS dataset is shown in the Figure 3.3.

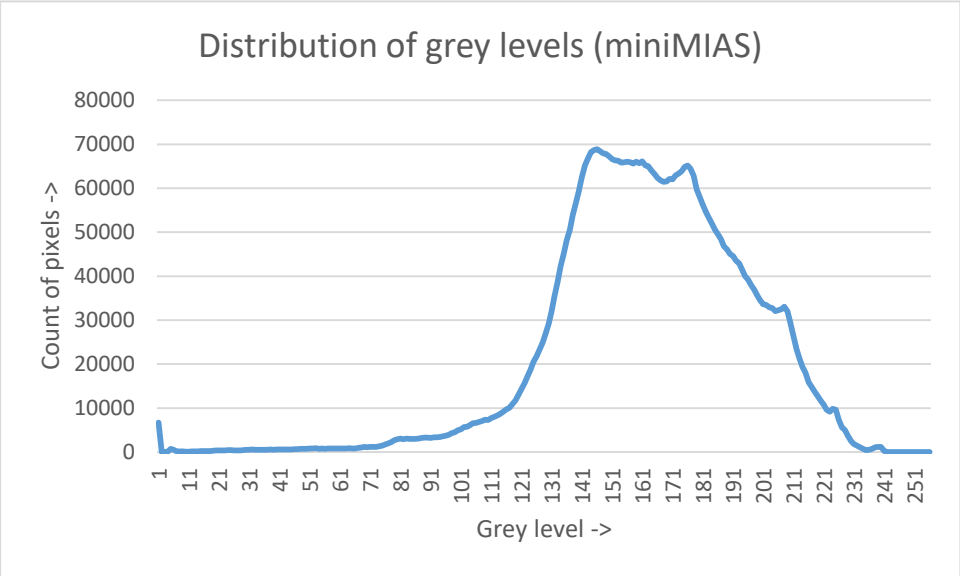


Figure 3.3 Distribution of grey levels in miniMIAS dataset

The database consist of 320 images out of which 270 images (24 circumscribed, 19 speculated and 207 normal) were used for the experiments. In this dataset the abnormality area or region of interest (ROI) is specified with the help of x, y image coordinates of centre of abnormality and the radius of a circle enclosing the abnormality. Information about the type of abnormality such as calcification, mass (circumscribed, speculated or ill-defined), architectural distortion or asymmetry is specified in the database. The type of background tissue such as Fatty, Fatty-glandular andDense-glandular has also been recorded for each mammogram. Sample ROI's extracted from miniMIAS dataset are shown in Figure 3.4 below.

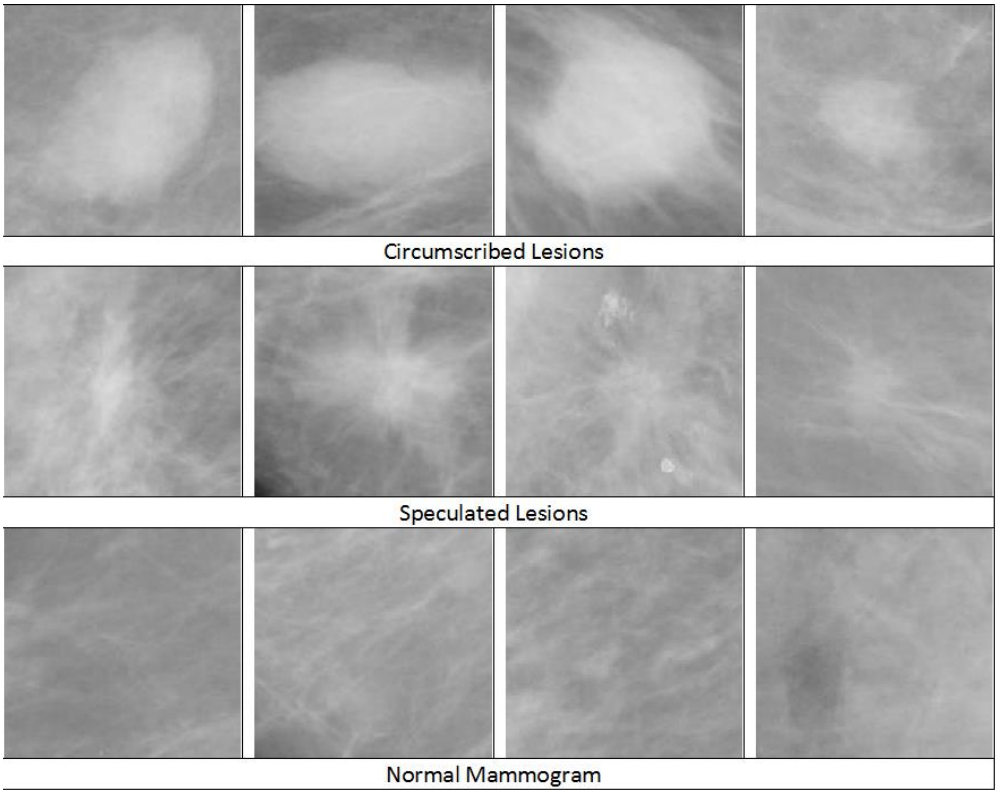


Figure 3.4 Sample ROIs from miniMIAS dataset

3.2.4 Lakeimaging dataset of mammography

LakeImaging Pvt. Ltd. Inc. is a diagnostic radiology provider in central and western Victoria, Australia. The dataset of annotated mammograms was specially prepared for this project. These mammograms are real word breast images collected from patients during the year 2013 and 2014. Images are de-identified and abnormalities are marked by experienced radiologists at LakeImaging. These images are available as presentation state Digital Imaging and Communications in Medicine (DICOM) objects. This dataset consists of 20 circumscribed mass, 20 speculated mass and 40 normal mammograms. Each abnormality was manually annotated describing shape, margin of the lesion, density of the background tissue, whether the lesion is benign or malignant and diagnosis of the lesion. This will serve as the ground truth for our classification experiments. These images are taken at various imaging stations and hence they are of different resolutions (1076 × 1586, 1532 × 1724, 2044 × 2236, 1636 ×1724). Some sample ROI images from LakeImaging database can be seen in Figure 3.5.

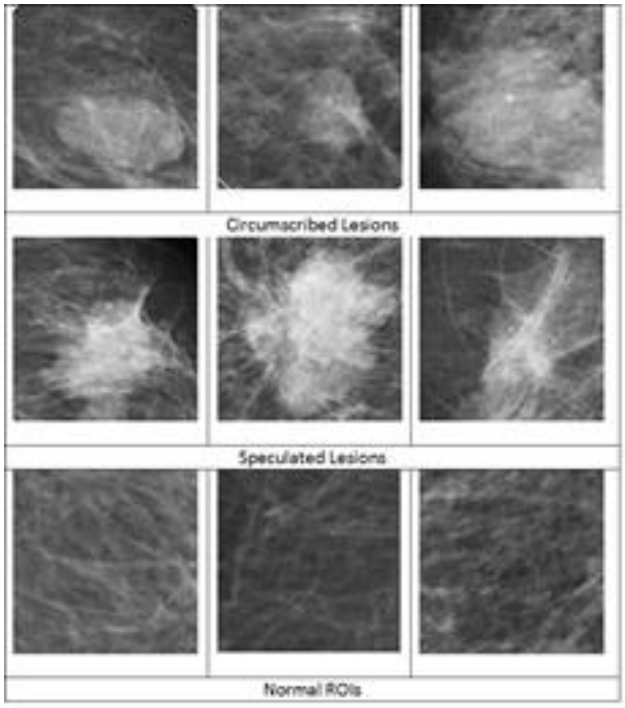


Figure 3.5 Sample ROIs from Lakeimaging dataset

3.3 Overall System Design

As stated earlier, the main aim of this research is to classify mammographic images. Two types of classifications have been tried in this study. First is the normal/abnormal classification which is useful for the automated detection (CADe)/screening process and second is the classification of ROIs into circumscribed, speculation and normal categories which is useful for the automated diagnosis (CADx) of breast cancer. A novel feature extraction technique Pixel N-grams inspired from the character N-gram concept in text categorisation domain has been proposed for the classification. This technique is explained in detail in the Section 3.3.2. Figure 3.6 illustrates a schematic overview of the approach deployed to apply Pixel N-grams to mammograms classification. Each stage is explained in the following subsections.

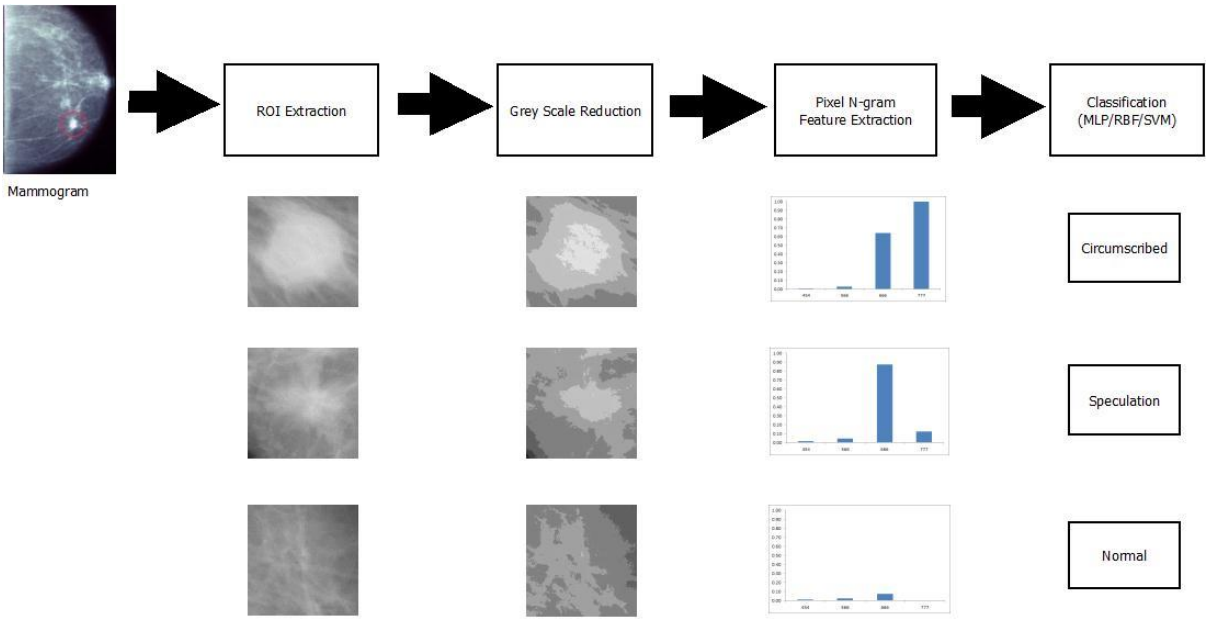


Figure 3.6 Schematic overview of experimental procedure

3.3.1 Pre-processing

The first step of mammographic image classification is pre-processing. The pre-processing step consists of downscaling, region of interest (ROI) extraction and grey scale reduction steps as explained below.

3.3.1.1 Downscaling

The mammographic images in Lakeimaging dataset are not of same resolution. In order to get equal resolution for all the images, the images were scaled to the lowest resolution image available in the dataset which was 1076 × 1586 pixels. There were two options to make all the images of equal resolution; either to downscale to the lowest resolution available in the dataset or to upscale all the images to the highest resolution available in the dataset. The upscaling of images even with sophisticated interpolation algorithms, has a high possibility of introducing noise. On the other hand downscaling can basically reduce some noise by removing redundant pixels and also make the feature extraction faster due to smaller number of pixels to analyse. Downscaling of images refers to resizing/resampling of the image to lower number of pixels. Although, downscaling could possibly result in some information loss, the advantages outweigh the disadvantages here. Therefore, downscaling of images was chosen for our experiments. To downscale the images accurately with minimal information loss or introduction of noise, the use of an interpolation technique was adopted.

Interpolation means using known data to estimate values at unknown points. Grey scale image interpolation works in two directions. It uses the intensities of the surrounding pixels and tries

to estimate a best value for a pixel's intensity. There are different rescaling algorithms such as nearest neighbour, bi-linear interpolation, bi-cubic interpolation, spline and sync. Bi-cubic interpolation considers the closest 4x4 neighbourhood of known pixels (16 pixels) to estimate the value of an unknown pixel. Depending on the distance of the known pixel from the unknown pixel, these intensity values are given different weightings in the calculation. Bi-cubic interpolation produces noticeably sharper images than nearest neighbour and bi-linear interpolation technique, and is the ideal combination of processing time and output quality (Doma, 2008). Therefore, we chose Bi-cubic interpolation technique for downscaling the images from LakeImaging dataset.

Images from mini-MIAS dataset are of 1024×1024 resolution and hence do not need downscaling.

3.3.1.2 ROI Extraction

The abnormalities/lesion are concentrated in highly localised areas in mammograms and these are called as Regions of Interest (ROI). The radiologists are basically interested in these areas. Therefore, it is necessary to extract ROIs and classify these in order for accurate diagnosis of the lesions. For both the datasets (miniMIAS as well as LakeImaging) we have the abnormalities marked by experienced radiologists.

In the miniMIAS mammography database, the abnormalities are described using x, y image co-ordinates of the centre of abnormality and the radius of a circle enclosing the abnormality. On the other hand Lakeimaging dataset consist of the DICOM presentation state objects/images

where the abnormalities we marked as circles by the experienced radiologist. Sample abnormality annotation for miniMIAS and Lakeimaging dataset is shown in Figure 3.7.

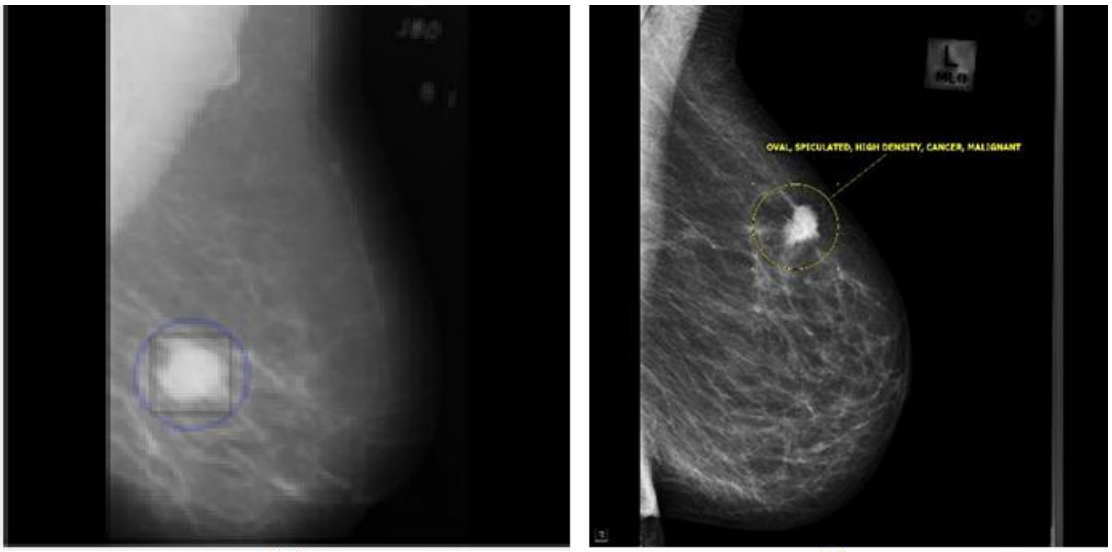


Figure 3.7 Annotated ROI's (a) miniMIAS (b) Lakeimaging

Equal sized square regions surrounding the abnormalities were cropped (after downscaling images for Lakeimaging dataset) using MATLAB software. These cropped images are called Regions of Interests (ROIs) and saved in the database. The ROI size of 140 × 140 pixels was decided as this size included all of the lesions completely. For normal mammograms the same size regions of interest are extracted from the centre of the mammogram.

3.3.1.3 Grey Scale Reduction

In the ROI images, each pixel is represented with a grey level between 1 and 256. One of the main goals of this research is to design a computationally efficient system for classification of mammograms. Drastic reduction in computational complexity can be achieved by reducing the images in Grey scales. Reducing images in grey scale also minimises the effect of noise. Various types of strategies for grey scale reduction have been tried. These strategies for grey scale reduction are explained below. Basically we are trying to discretise the grey level pixel data into a smaller range so as to reduce the feature computation cost. There are mainly two strategies for discretisation of data namely equal size binning and equal frequency binning (S. Kotsiantis & Kanellopoulos, 2006).

3.3.1.3.1 Equal Size Binning

Combining range of intensity values and treating as one intensity is called as binning. If bins are created such that all bins have equal number of intensity values, it is called as equal size binning. In this case the bin size can be calculated by equation 3.7, where G_H = Highest grey level, G_L = Lowest grey level, N_B = Number of grey level bins.

$$binsize = \frac{G_H - G_L}{N_B} \tag{3.7}$$

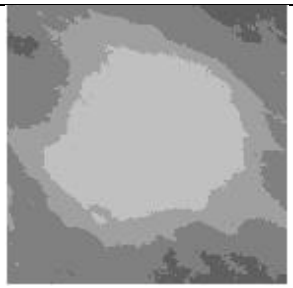
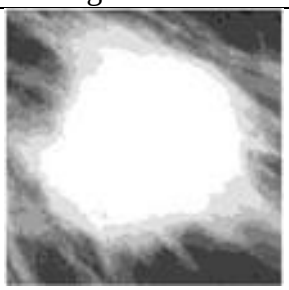
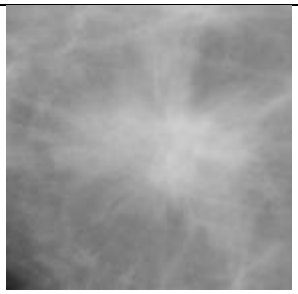
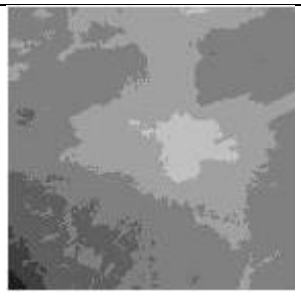
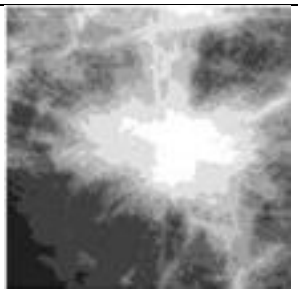
And thus i^{th} bin is composed of intensity levels between I_{start} to I_{end} and are given by equations 3.8 and 3.9.

$$I_{start} = (i - 1) \times binsize \tag{3.8}$$

$$I_{end} = (i \times binsize) - 1 \tag{3.9}$$

3.3.1.3.2 Equal Frequency Binning

In equal frequency binning, the start and the end grey level of the bin is decided so that each bin contains approximately the same number of pixels. Figure 3.8 shows sample images from miniMIAS dataset reduced in grey scale using equal size binning and equal frequency binning.

Lesion type	Original ROI	Equal size binning	Equal frequency binning
Circumscribed			
Speculated			

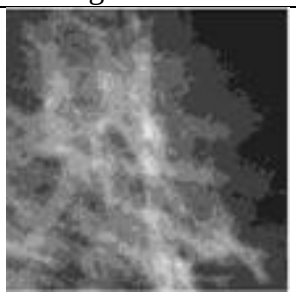
Lesion type	Original ROI	Equal size binning	Equal frequency binning
Normal			

Figure 3.8 Grey scale reduction using equal size binning and equal frequency binning

3.3.2 Pixel N-grams feature extraction

After grey scale reduction of the ROI images, the next step is to calculate the Pixel N-gram features of an image. As stated earlier in Chapter 2, Pixel N-grams are inspired from character N-grams for text categorisation. Here grey scale intensity of a pixel is equivalent to one visual character. Pixel N-grams are the sequence of N consecutive grey level pixels in an image. This technique is therefore called visual character N-gram/Pixel N-gram model. One gram features would thus be similar to the histograms of each grey levels present in an image. Two gram features consider every possible combination of the two grey levels adjacent to each other. This can be closely related to co-occurrence matrix features with an adjacency distance of 1.

However, the application of Pixel N-gram model for images is not straightforward. While text documents have a single spatial direction, images are two dimensional and the sequence of feature descriptors can be obtained in different orientations (vertical, horizontal or at an angle of Θ degrees). Figure 3.9 shows the sliding window for calculating three-gram features (image reduced using 8 grey level bins) in a horizontal and vertical direction.

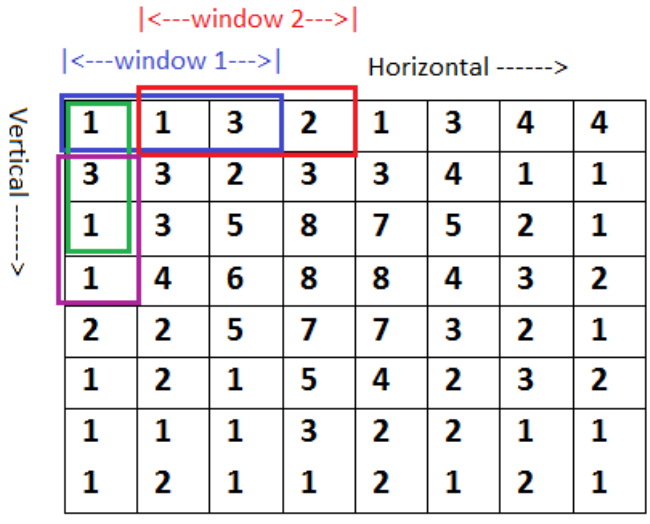


Figure 3.9 Sliding window for 3-gram computation

The computational complexity and feature vector dimension increases rapidly with increase in N. Apart from resulting in large vocabularies, longer sequences are harder to find and hence create sparse feature vector. Therefore, the value of N has been restricted to 1, 2, 3, 4 and 5 for our experiments.

Figure 3.10 describes the process of Pixel N-gram representation of ROIs. It is achieved in two steps. The first step is the creation of the Pixel N-gram/visual character N-gram codebooks. All the images are considered during this step. The grey scale reduction of ROIs is achieved as explained in the aforementioned Section 3.3.1.3. The number of possible N-grams is dependent on the grey level bins used to represent the image and can be calculated with equation 3.10 where, N_p = Number of possible N-grams, N_g = Number of grey levels and N= Number of adjacent pixels considered for computing N-grams.

$$N_p = (N_g)^N \tag{3.10}$$

For example, possible number of 3-grams with ROI grey scale reduction to 8 grey levels are 8^3 = 512, number of possible 4-grams are 8^4 = 4096. Moreover, all the possible N-grams are not necessarily present in the given corpus producing zero counts for the N-grams absent in the corpora. This creates a sparse vector problem. The sparse vector problem can be dealt with using the N-grams present in the corpus by generating a codebook/dictionary.

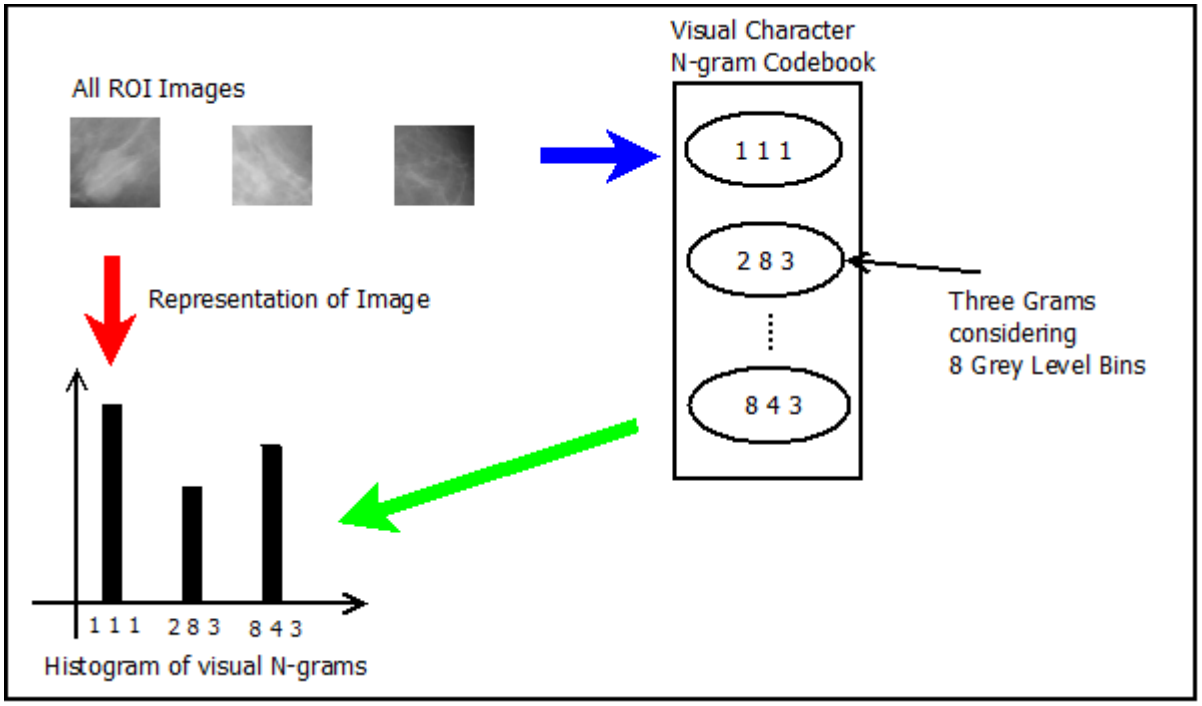


Figure 3.10 Pixel N-gram representation of ROIs

For the experimental analysis the N-grams in horizontal and vertical directions were considered. Although, N-grams having the same order but different orientations convey totally different meanings in case of text, for an image they may correspond to the same pattern.

Figure 3.11 shows the horizontal and vertical 3-grams for a 3×3 pixels image after rotation and flipping. Consider if the image is rotated by 90 degrees then the pattern 4-3-2 explains the distribution of the same pixels which generated the 2-3-4 pattern. Thus in order to have rotation invariant N-gram features both these patterns are considered the same N-gram.

		Horizontal 3-grams	Vertical 3-grams
Original Image	2 8 3 3 3 4 4 6 8	2 8 3 3 3 4 4 6 8	2 3 4 8 3 6 3 4 8
90 degree rotated image	4 3 2 6 3 8 8 4 3	4 3 2 6 3 8 8 4 3	4 6 8 3 3 4 2 8 3
Flipped Image	3 8 2 4 3 3 8 6 4	3 8 2 4 3 3 4 6 4	3 4 8 4 3 6 2 3 4

Figure 3.11 Effect of image rotation on Horizontal and vertical N-grams

In the second step using the Pixel N-gram codebooks, histogram of occurrence of N-grams is calculated for every ROI.

3.3.3 Feature Normalization

A typical classifier algorithm performs better when the input data is within a standard range (Aksoy & Haralick, 2001). Transforming the input data within the required range is known as normalization/scaling. There are various ways to do this type of transformations. Normalization or scaling the features can make training of the classifier faster by reducing the chances of getting stuck in a local minima. Different types of widely used normalisation strategies are explained below.

In Zscore normalisation features are rescaled so that they have the properties of a standard normal distribution. It is calculated with the Equation 3.11, where, z = normalised value of the feature, x = feature value, μ = mean and σ is standard deviation.

$$z = \frac{x - \mu}{\sigma} \tag{3.11}$$

Another approach for normalisation is Min-Max scaling. Here, the data is scaled to a fixed range usually 0-1. This way smaller standard deviations can be achieved which suppress the effect of outliers. It is calculated with the Equation 3.12 where, x_{norm} = normalised value of feature, x_{min} = minimum value of feature and x_{max} = maximum value of feature in the corpus.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.12}$$

From information theory it is clear that some of the features have a little discriminating power in determining relevance (Robertson, 2004). Therefore, choosing a term capable of discriminating between various classes definitely increases the classification performance. For example, commonly appearing words such as 'and', 'the', 'an' do not play any important role in classification. Hence it is necessary to weigh the features using different weights.

The most commonly used weighting strategy is tf-idf weighting where tf is the frequency of appearance of word in a document and idf is the inverse document frequency and is calculated by using the Equation 3.13 where, N = total number of documents and N_t = number of documents that contain the word. Idf was proposed with an exploratory intuition that a term occurring in many documents is not a good discriminator and should be weighted less than the one which occurs in few documents (Spärck Jones, 2004). Thus idf is the measure of how important a word is in the given corpus.

$$idf = log \frac{N}{N_t} \tag{3.13}$$

Thus more weight can be given to the more important features in the dataset. A similar analogy is true for the images. In order to achieve high classification performance these three types of normalization techniques are tried in the experiments.

3.3.4 Classification

After normalisation of features, the next step is classification of the ROIs based on the normalized N-gram counts. The confusion matrix from the classification is noted and the performance is compared using the predictive accuracy, sensitivity/recall, specificity, precision, TP rate, FP rate and Fscore.

Many different classifiers have been used for image classification tasks (S. B. Kotsiantis, Zaharakis, & Pintelas, 2007). Each has advantages and limitations. Three most commonly used

classifiers (Multilayer Perceptron - MLP, Support Vector Machine - SVM and K-nearest Neighbour - KNN) were used for the experiments. Basically, the MLP classifiers are good with noise and uncertainty and consider all the input features for determining the output class. On the other hand SVM classifiers try various combinations of subset of input feature-values to find out the best feature subset to separate out two classes. Thus it ignores many features but could be useful to determine if there are any prominent N-gram features which are able to distinguish among the various lesion classes well. KNN classifier is the simplest yet powerful classifier which is generic, can handle many classes and has worked for variety of datasets. Each of these three classifier functionality, advantages and disadvantages are detailed below.

A Multilayer perceptron is a feed-forward artificial neural network where neurons of i^{th} layer serve as input features for neurons of $(i + 1)^{th}$ layer. The MLP classifiers are parameterized non-linear models in which weights of each layer are adjusted such that the sum of square error between the predicted output and the expected output is minimized. It uses supervised learning rule called as backpropagation. Figure 3.12 shows a backpropagation multilayer perceptron with two hidden layers. The weights of connections between the neurons of two different layers are adjusted such as to get a minimum error between the actual outputs and expected output values.

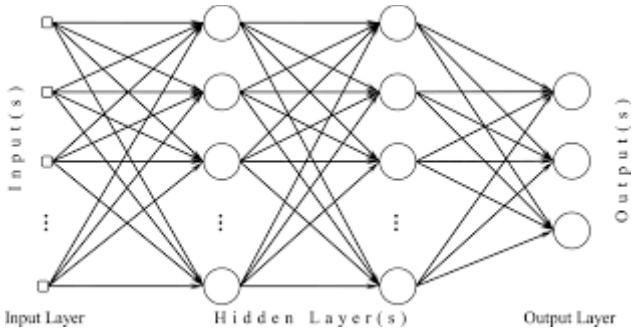


Figure 3.12 Classification using MLP classifier

According to Bengio (Bengio & LeCun, 2007) deep architectures such as MLP can represent functions more accurately than shallow architectures such as SVM. However, determining the size of hidden layer is a problem as few neurons lead to poor approximation and generalization capabilities whereas, higher number of neurons can result in overfitting the space and considering noise in the input data (S. B. Kotsiantis et al., 2007).

Support Vector Machine (SVM) looks for a hyperplane that separates one class from the other. Let us consider if we have data points belonging to two classes as shown in the Figure 3.13 and

goal is to classify a point to either of these classes. There are many possible hyperplanes which can separate the two classes. A hyperplane where distance from it to the nearest data point on each side is maximized is called as optimal hyperplane.

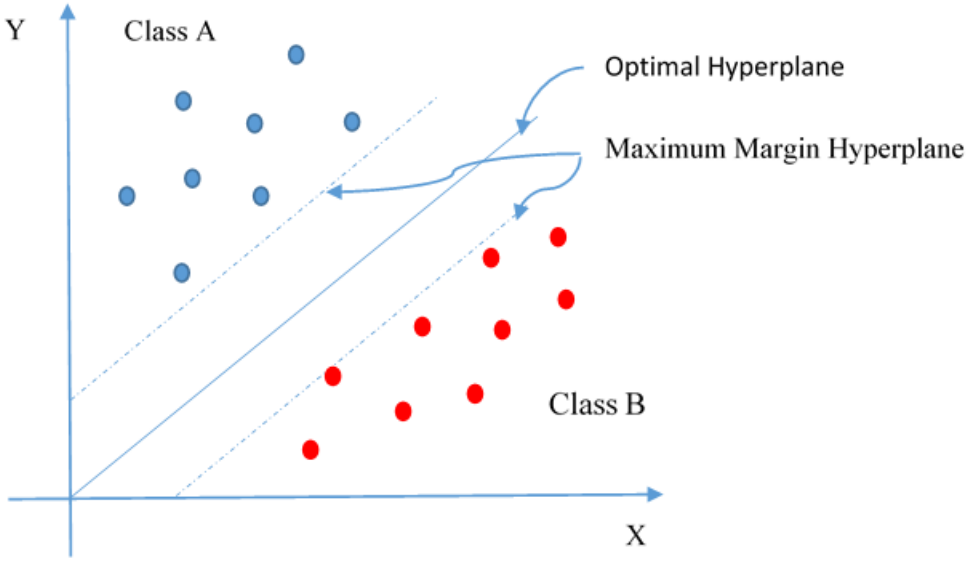


Figure 3.13 Classification using SVM Classifier

These classifiers are seen as non-parametric linear models where training is achieved by explicit determination of the decision boundaries from the training data.

K-nearest neighbour (KNN) is a non-parametric, flexible, simplest machine learning algorithm used for classification applications. To classify an unknown instance, the distance from that instance to every other training instance is measured. Output class label is decided by identifying the k closest instances and finding the most frequently represented class in these k classes. Figure 3.14 shows how a KNN classifier decides the class of an unknown example. Generally an odd number is used for the parameter k to ensure no tie exists.

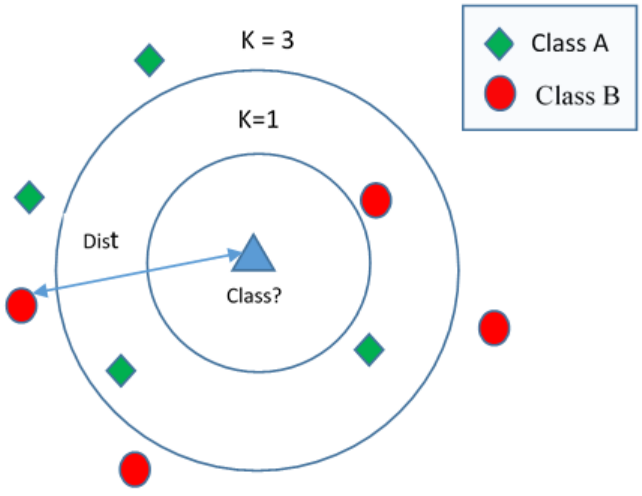


Figure 3.14 Classification using KNN classifier

When percentage of training samples for one class is same as that of the other class, the training is said to be balanced learning. However, in many of the datasets this is not the case. It has been observed that ANN outperforms Support Vector Machine (SVM) when balanced learning is absent and the performance of both the classifiers become comparable with the balanced learning (Ren, 2012). Classifiers such as MLP and SVM requires intensive learning/training stage. On the other hand KNNs can naturally handle many number of classes, avoid overfitting, and most importantly require no training phase (Boiman, Shechtman, & Irani, 2008). The training time required is zero in this case. However, as KNN must compute the nearest neighbours for each sample during the classification phase, more time is required during the classification stage. Thus KNN employs instance based or lazy learning strategy where function is approximated locally, and hence the prediction stage is very slow (Stopel, Boger, Moskovitch, Shahar, & Elovici, 2006). Further, usually quantization and informative feature selection is used for dimensionality reduction of the feature vector before the image classification task. This incurs large information loss. It has been observed that the quantisation thus degrades the performance of non-parametric classifiers such as KNN (Boiman et al., 2008).

3.4 Experimental Methodology

In order to fulfil the research aims various experiments have been designed. The following experiments were carried out:

1) Finding optimum value of N

2) Comparison of different classifiers

3) Effect of choosing different normalisation techniques

4) Comparison with existing techniques (classification performance)

5) Comparison with existing techniques (computational complexity)

6) Feature selection using wrapper approach

7) Piecewise constant approximation

8) Size and location invariance for shape classification

9) Resolution invariance for shape classification

These experiments are described below in detail.

3.4.1 Finding optimum value of N

The value of N is expected to have a significant effect on the classification performance. As N is increased, more details regarding the spatial relationships among the pixels are incorporated into the Pixel N-gram features and hence better representation of image can be expected. However, with increase in value of N, the Pixel N-gram image features would become more specific to a particular image making it hard for a classifier to generalize.

Thus, it can be hypothesized that the classification performance would increase as N is increased up to certain level but most probably start decreasing with further increase in N. The value of N where the maximum classification performance can be achieved would be called as optimum value of N.

The objective of this experiment is to find out this optimum value of N. This optimum value would probably be dependent upon the database under consideration as the distribution of grey levels in each database differs. Another aim of this experiment is to verify if the optimum value of N is dependent upon the database under consideration. Therefore, the experiment was conducted on 4 different datasets namely miniMIAS, LakeImaging, UIUC texture dataset and basic shapes dataset. The images in the miniMIAS, LakeImaging and UIUC texture dataset have grey scales ranging from 1 to 256 whereas, the binary shapes database only contains two grey levels (0 and 1).

All the images from LakeImaging dataset are not of same resolution hence they are downsized using bi-cubic interpolation technique to the lowest resolution available in the dataset (Refer to Section 3.3.1.1). On the other hand images from miniMIAS dataset, UIUC texture dataset and basic shapes dataset are of same size and resolution and hence do not require downsizing. The equal sized (140 × 140 pixels) regions of interests (ROI) were then extracted from the images.

Low computational cost is necessary for increasing the efficiency of the radiologists. Therefore, ROI images were grey scale reduced using 8 grey levels for reducing the computational cost. The best value of grey scale reduction to 8 was decided by analysing the number of actual N-gram features and possible number of N-gram features in the corpus. This

analysis is detailed in the Chapter 5. It is surmised that the grey scale reduction would reduce the noise level by ignoring redundant pixels. Both the binning strategies, equal size binning and equal frequency binning were tried for reducing the images in grey scale in order to find out which one gives best classification performance.

N-gram features were computed using the method explained in Section 3.3.2. Min-max normalization is the simplest normalization technique which preserves the relationships among the original data values. Therefore, N-gram features were normalised using min-max normalisation (Refer to Section 3.3.3). The classification of lesions into three categories (circumscribed, speculation and normal) was observed using MLP classifier. To find the optimum value of N classification using N= 1, 2, 3, 4 and 5 were considered.

Generalisation was estimated using 10 fold cross-validation resampling. Classification performance was measured using Fscore, Sensitivity and Specificity metrics.

3.4.2 Comparison of different classifiers

The objective of this experiment was to compare different classifiers in order to find out which one works best for mammographic classification. All of the N-gram features seem to be important for distinguishing one class from another. Therefore, it was hypothesized that the MLP classifier would work better than the SVM and KNN classifiers for the mammographic lesion classification. The experiment was carried out on the miniMIAS, LakeImaging as well as shapes dataset.

The best value of grey scale reduction and optimum value of N obtained from the experiment described in section 3.4.1 was used for reducing the ROIs in grey scale and computing N-gram features. The N-gram features were normalised using the min-max normalisation explained in Section 3.3.3. The normalised N-gram features were fed as inputs to the classifier.

As discussed in Section 3.3.4 every classifier has some advantages and limitations. The most commonly used classifiers for image classification are MLP, SVM and KNN. These three classifiers were used in order to determine which classifier performs best for the mammograms classification task. 10-fold cross-validation technique was used for estimating the generalisation. Performance was compared using Fscore, Sensitivity and Specificity metrics.

3.4.3 Effect of choosing different normalisation techniques

For this experiment, three widely used normalisation techniques (zscore, min-max and tf-idf) were employed before classification. For more details please refer to Section 3.3.3. This experiment was conducted on miniMIAS dataset of mammography as it is a benchmark database, contains good number of images and annotated by many radiologists.

The ROIs were extracted, and grey scale reduced to 8 grey levels. N-gram features were computed considering the optimum level of N obtained from experiment described in section 3.4.1. Then the N-gram features were normalised using the three above mentioned techniques. Finally, the classifier with optimum performance, chosen from experiment described in section 3.4.2 was considered for classification. The classification using different normalisation techniques was compared using classification Accuracy, Sensitivity and Specificity values.

3.4.4 Comparison with existing techniques (Classification Performance)

In order to compare the classification performance using Pixel N-gram features, various widely used existing techniques were considered. First technique used for comparison was intensity histogram. Second most prevalent method for extracting texture features for mammographic classification was Grey Level Co-occurrence Matrix (GLCM) based Haralick's features (Haralick et al., 1973).

The experiment was conducted on miniMIAS as well as LakeImaging dataset. Images from LakeImaging dataset were downsized to the lowest resolution image in the dataset in order to maintain consistency. Images from miniMIAS dataset are of same resolution. The ROIs are extracted by cropping an equal sized region (140×140) surrounding the abnormality. Then the Pixel N-gram features were computed considering the optimum value of N. These features were then normalised using best normalisation technique obtained from the results of the experiment described in Section 3.4.3. The normalised features were then used as input to the classifier. The classifier having high performance (Refer to experiment described in Section 3.4.2) was considered here for classification.

Two types of classification were observed. One is normal/abnormal classification. The second is circumscribed/speculation/normal classification. Generalisation was estimated using 10-fold cross-validation resampling for miniMIAS dataset due to good number of images. On the other hand leave-one-out cross-validation was employed for estimating the generalisation for Lakeimaging dataset due to less number of images. The performance was compared using Fscore, Sensitivity and Specificity values.

Comparison of classification using Pixel N-grams is also performed using UIUC texture dataset. For the UIUC dataset, the whole image was considered as ROI and is grey scale reduced using optimum value of grey level. Then the N-gram features were computed using the optimum value of N for the texture dataset. The min-max normalisation was used for transforming the N-gram counts in the range of 0-1. Then the SVM classifier was used for classifying the texture images into 25 different classes. The classification performance was compared against the classification using an intensity histogram, Haralick's features as well as BoVW method. The performance was compared using Fscore, Precision and Recall measures.

Classification using Pixel N-gram features was also compared using the specially prepared shapes dataset. The images in this dataset were binary images and hence there were only two grey levels 0 or 1. The classification of shapes in three categories (circle, square, triangle) was observed using MLP classifier. The performance was compared with intensity histogram as well as Haralick's features.

For calculating histogram features, co-occurrence matrix based Haralick's features and Pixel N-gram features, algorithms were implemented using Matlab software.

3.4.5 Comparison with existing techniques (Computational Complexity)

In mammographic classification, computational cost is one of the important factors needed for increasing the efficiency of the radiologists. In order to exploit the computational efficiency of the Pixel N-gram features, the time required to compute the 3-gram features was compared with the time required for computing various other widely used features such as Intensity histogram and Haralick's features.

This experiment was conducted on miniMIAS dataset due to good number of images in the dataset. In order to analyse how the computational complexity increases with increase in the image size in pixels, ROIs of four different sizes (70×70, 140×140, 280×280, 560×560) were extracted from every mammogram image. The time required to compute 3-gram features in horizontal and vertical directions was noted for all the ROIs. Then the average time required for four different image sizes is calculated and noted. Similarly, the average time taken for computation of intensity histogram features and Haralick features was noted for every ROI. For computation time measurement, standard Tic and Toc functions available in Matlab were used.

3.4.6 Feature selection using wrapper approach

In classification, most of the times it is possible that some features are redundant or irrelevant. The redundant features cause more training time whereas, the irrelevant features can result in overfitting and hence less generalised classification model. The process for identifying the best subset of features from all the features is known as feature selection. Feature selection is performed for three main reasons.

- Reduce overfitting and hence better generalisation
- Improve accuracy
- Reduce training time

The objective of this experiment is to find if a subset of features exists which can distinguish between the circumscribed/speculation/normal classes and to determine the most significant features in order to reduce the training time and improve classification accuracy.

Weka datamining software (Weka, 2011) provides feature selection tool using the attribute evaluator and search method. The most commonly used feature selection methods are a) wrapper method b) filter method. Wrapper method (wrapper subset evaluator, weka) has been used here to find the most significant features. A wrapper subset evaluator creates all possible subsets of features from the feature vector. Then it uses the classifier algorithm using each subset. The subset of features with highest classification performance is declared as the best feature subset. The 3-gram features in horizontal and vertical direction are considered and the subset is decided using the wrapper subset evaluator with a MLP classifier.

3.4.7 Piecewise constant approximation

In the experiments conducted so far, value of grey scale for grey scale reduction was decided by analysing the possible and actual number of Pixel N-grams and empirically by trying out different values. The empirical task is not only time consuming but may not be accurate as it is almost impossible to try out each and every possible value of grey scale. Further, the grey scale reduction required may quite possibly be different for different datasets and different classification problems. Thus, in order to achieve high classification performance without having to perform grey scale reduction, a piecewise constant approximation method could be used.

For this method N-gram features of every image were computed considering all the 256 grey levels. Thus a list of points (N-grams) and their associated function values (number of occurrences/counts) were obtained. The idea is then to approximate these values with a simpler function for each image. The simpler function can be computed by using piecewise constant approximation. Piecewise constant functions are functions for which the space can be subdivided into subspaces over each of which the function takes the same value. This would result in a reduced number of parameters or sets of coefficients which could then be used as features for classification of images.

3.4.8 Size and location invariance for shape classification

Classification of shapes using Pixel N-grams was observed in this experiment. The basic shapes dataset explained in Section 3.2.2 was used.

In the first experiment, 3 basic shapes of 20 different sizes each (60 images altogether) were considered. Pixel N-gram features were calculated for every image in the dataset. These features were then normalised using min-max normalisation and used as input to the MLP classifier for classification into 3 classes (circle, triangle and square). The results were compared with Intensity histogram and Haralick's features. For comparison purpose intensity histogram features were computed for the same dataset and classification using MLP classifier was conducted. Similarly, four main Haralick's features based in co-occurrence matrix (contrast, correlation, energy and homogeneity) were computed. These features were normalized and used for classification using MLP classifier. Due to small dataset size, the generalisation was estimated using leave-one-out validation resampling.

The objective of the second experiment was to see how accurately N-gram features can classify different shapes located at different sites in an image. For this experiment, various images were created by changing the shape locations. Basic shapes dataset of images of same resolution consisting of 80 circles, 80 triangles and 80 squares was used for this experiments.

The optimum value of N obtained from the experiment described in Section 3.4.1 was considered. N-gram features in horizontal as well as vertical directions were computed and given as input to the MLP classifier. Leave one out classification was used to estimate the generalisation. Performance was noted using classification accuracy for each class.

3.4.9 Resolution invariance for shape classification

This experiment was carried out in order to discover to what extent classification using Pixel N-gram features is resolution independent.

10 different resolutions of three basic shape images were considered for this experiment. The 3-gram features were found to be effective for shape classification according to the experiment explained in section 3.4.1 and hence 3-gram features in horizontal and vertical directions were computed. These N-gram features were normalized using min-max normalisation and were used as input to the MLP classifier for classification of shapes into circle, triangle and square categories. The Fscore, precision and recall values were noted using leave one out validation. The same experiment was carried out using Intensity histogram features. Similarly, classification was also achieved using Haralick's features. The results were then compared and two tailed paired T-test was conducted to see if the results using Pixel N-gram features were significantly better than the other two techniques.

3.5 Chapter Summary

In this chapter the research approach including performance criteria for evaluation is outlined. The four datasets used for the experiments are then explained. Various steps necessary for classification of images using Pixel N-gram technique are then described. The rationale and the procedure for conducting various experiments were then detailed. The experimental results on texture and shape datasets are described and analysed in Chapter 4 and the experimental results on mammographic image datasets (miniMIAS and LakeImaging) are noted and analysed in Chapter 5.

4 Experimental Results and Analysis (Texture and Shape)

This chapter describes the classification experiments using Pixel N-gram features conducted on texture and shape images. As per BIRADS standard (Radiology, 2013) mammographic lesions are characterised by texture and shape. Therefore, it becomes necessary to explore to what extent Pixel N-gram features can result in classifiers that distinguish between various textures and shapes.

Experiments were conducted on the benchmark texture image database (UIUC) (http://www-cvr.ai.uiuc.edu/ponce_grp) to explore the use of Pixel N-gram features for texture classification. Section 4.1 details these experiments. These experiments were primarily designed to compare the performance of Pixel N-gram features with existing feature extraction techniques such as Intensity histogram, Haralick's features and BoVW features.

Lesions in mammographic images are of different shapes and sizes. They could be located at different sites in an image. Above all, these mammograms produced at different imaging stations could be of different resolutions. Motivated by these properties of mammographic images, it becomes important to evaluate the performance of Pixel N-gram features for classification of lesions considering various shapes, shape sizes, shape locations and different resolutions of images. The use of Pixel N-grams for the shape classification (Refer to Section 4.2) was assessed with a database of binary images (black and white) specially prepared with three basic shapes (circle, triangle and square). The experiments on shapes dataset are designed to analyse the performance of Pixel N-grams for shape classification and to demonstrate that these features are size, location and resolution invariant.

The datasets used and the methodology for the experiments are detailed in Chapter 3.

4.1 Experiments on Texture Dataset

This research project is mainly focused on detection and diagnosis of breast mass lesions. Breast mass are cells that are more dense than the surrounding tissues. The most important information useful for determining the probability of mass lesion being benign or malignant is features such as size, shape and location of the lesions. However, there are some lesions without a well-defined boundary that hinder a correct visualization. Interpretation of these types of lesions can result in number of false positives and hence increase in number of biopsies. Another important property of mass lesions useful for determination of malignancy is texture (da Rocha et al., 2016). Thus mammogram images that do not present well-defined contours

can be detected using texture features, thereby providing the expert with greater support in the diagnosis of breast cancer.

Texture is a feature that is difficult for humans to analyse. It is defined as quantitative measure of arrangement of intensities in an image and can be modelled using statistical, spectral or structural approaches (J. Zhang & Tan, 2002). The purpose of these experiments is to apply the Pixel N-gram representation of images for classification of grey scale texture images (UIUC dataset) and compare the performance with the texture classification using existing low level techniques such as histogram, statistical techniques such as co-occurrence matrix and state of the art Bag-of-Visual-Words (BoVW) representation performed by (Lazebnik et al., 2005).

The results of the experiments on UIUC texture dataset were published in one of our conference papers (P. Kulkarni, Stranieri, A., Ugon, J., Aug 2016).

4.1.1 Finding optimum value of N

The objective of this experiment is to use Pixel N-grams for the classification of texture images and analysis of the effect of varying N on classification performance. The value of N = 1, 2, 3, 4 and 5 were considered. N-gram features were computed and normalized using min-max normalization (Myatt, 2007). Classification was performed using the SVM classifier in WEKA datamining (Weka, 2011) software and generalisation was estimated using 10 fold cross validation resampling. Fscore for different classes are noted in the Table 4.1 below.

Table 4.1 Effect of varying N on texture image classification

Class	**1-gram**	**2-gram**	**3-gram**	**4-gram**	**5-gram**
bark1	0.345	0.609	0.699	0.789	0.729
bark2	0.633	0.720	0.785	0.867	0.835
bark3	0.765	0.765	0.880	0.935	0.910
wood1	0.114	0.469	0.805	0.824	0.802
wood2	0.393	0.857	0.842	0.921	0.885
wood3	0.581	0.642	0.750	0.815	0.772
Water	0.587	0.456	0.914	0.974	0.932
Granite	0.455	0.608	0.644	0.828	0.724
Marble	0.267	0.480	0.552	0.776	0.626
floor1	0.103	0.613	0.884	0.918	0.892
floor2	0.637	0.741	0.845	0.935	0.911

Class	**1-gram**	**2-gram**	**3-gram**	**4-gram**	**5-gram**
Pebbles	0.467	0.650	0.805	0.961	0.924
Wall	0.787	0.822	0.860	0.950	0.903
brick1	0.552	0.645	0.840	0.895	0.862
brick2	0.049	0.500	0.829	0.949	0.889
glass1	0.529	0.684	0.857	0.911	0.892
glass2	0.551	0.731	0.964	0.964	0.940
carpet1	0.747	0.963	0.967	0.976	0.962
carpet2	0.386	0.321	0.759	0.911	0.852
upholstery	0.964	0.988	1.000	1.000	0.988
wallpaper	0.175	0.438	0.539	0.747	0.689
Fur	0.468	0.460	0.611	0.789	0.752
Knit	0.161	0.277	0.592	0.779	0.754
corduroy	0.964	0.916	0.987	0.988	0.973
Plaid	0.933	0.949	0.975	0.987	0.975
Overall acc.	53.7%	66.1%	80.3%	**89.5%**	85.4%

It was observed that all the possible N-grams were not present in the given texture corpus. The number of N-grams actually present in the corpus with increase in number of N are given in the Table 4.2 below.

Table 4.2 Possible and observed N-grams for texture dataset

N	Possible N-grams	Observed N-grams	Percentage of possible N-grams observed
1	8	8	100
2	64	36	56.25
3	512	107	20.8
4	4096	654	15.9

The number of possible N grams is dependent on the grey level. At a grey level of 8, there are 8^1 1-grams and 8^2 2-grams and 8^n N-grams. Figure 4.1 illustrates that the number of N-grams observed in the texture database increases rapidly with an increase in N. Correspondingly, the percentage of possible N-grams that are observed decreases.

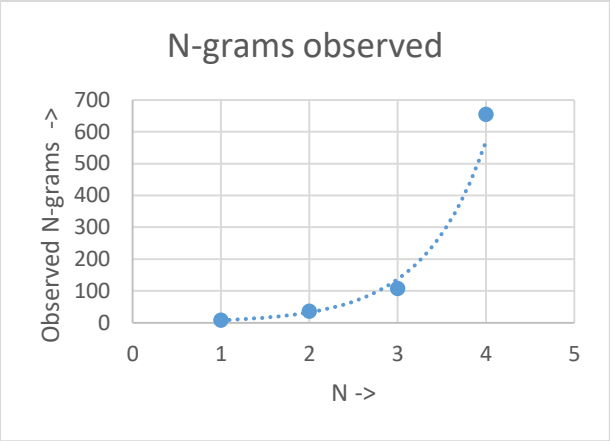


Figure 4.1 Number of observed N-grams with respect to increase in N

The number of observed N-grams can be modelled using the Equation 4.1 where N= number of consecutive pixels considered while computing N-gram features. Thus, as N is increased the percentage of the possible N-grams that are observed decrease exponentially. The classifier is trained on the observed number of N-grams. It is obvious that the classifier can generalise well if the N-grams used for training the classifier represent higher percentage of the possible N-grams which is at smaller values of N. However, N-gram features are able to model the image well with higher values of N at the cost of higher computational time. Thus there is always a trade-off between the good classification performance and computational complexity.

$$N_{observed} = 1.8769\, e^{1.43N} \tag{4.1}$$

Effect of varying N can be analysed using the graph shown in Figure 4.2. As N is increased the overall classification accuracy/Fscore is increased until N= 4 (89.5% for 4-grams). This could easily be described by the fact that as N is increased, more and more spatial relationships among the pixels are incorporated and the representation of the image becomes complete. However, the classification performance drops as the N becomes greater than 4. This could be because longer sequences become hard to find producing large vocabularies. Moreover, the resulting feature vector becomes high dimensional and contains image specific features. Therefore, these features with higher value of N tend to confuse the classifier. Thus N=4 can be seen as optimum value of N for this texture dataset. Furthermore, for almost every class in the texture dataset, classification performance becomes steady and independent of the texture classes with an increase in N. Also, with the increase in the value of N, the computational cost is increased. Therefore, a trade-off between the required classification performance and the computational overhead needs to be considered while deciding N for a particular application.

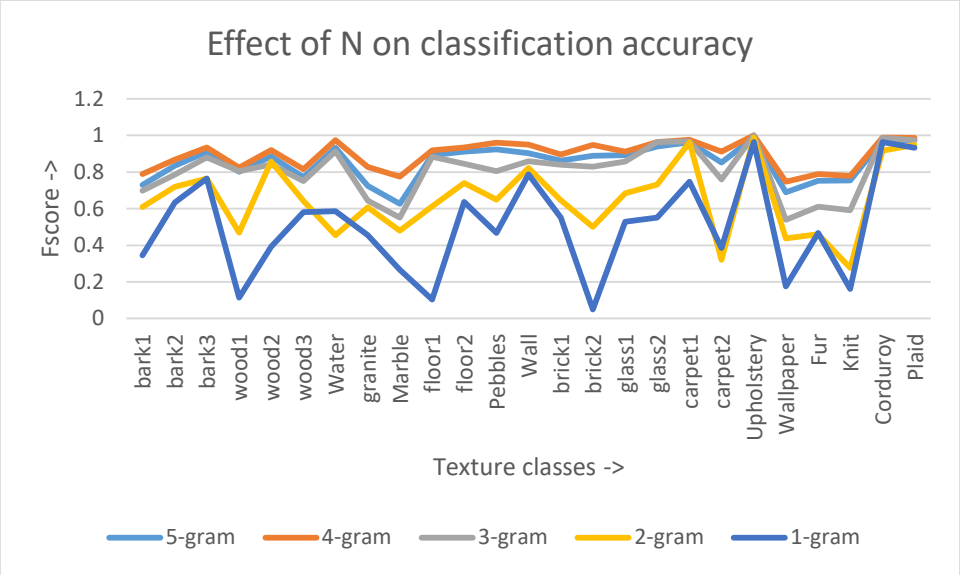


Figure 4.2 Effect of varying N on classification accuracy

Further, it was observed that for the classes such as carpet1, upholstery, corduroy and plaid there is an almost negligible effect of increasing N on the classification accuracy. It is observed that these images have regular patterns (distributed evenly) throughout the image (Refer to Figure 4.3). Also, textures seem to be equally illuminated on all the parts of the image.

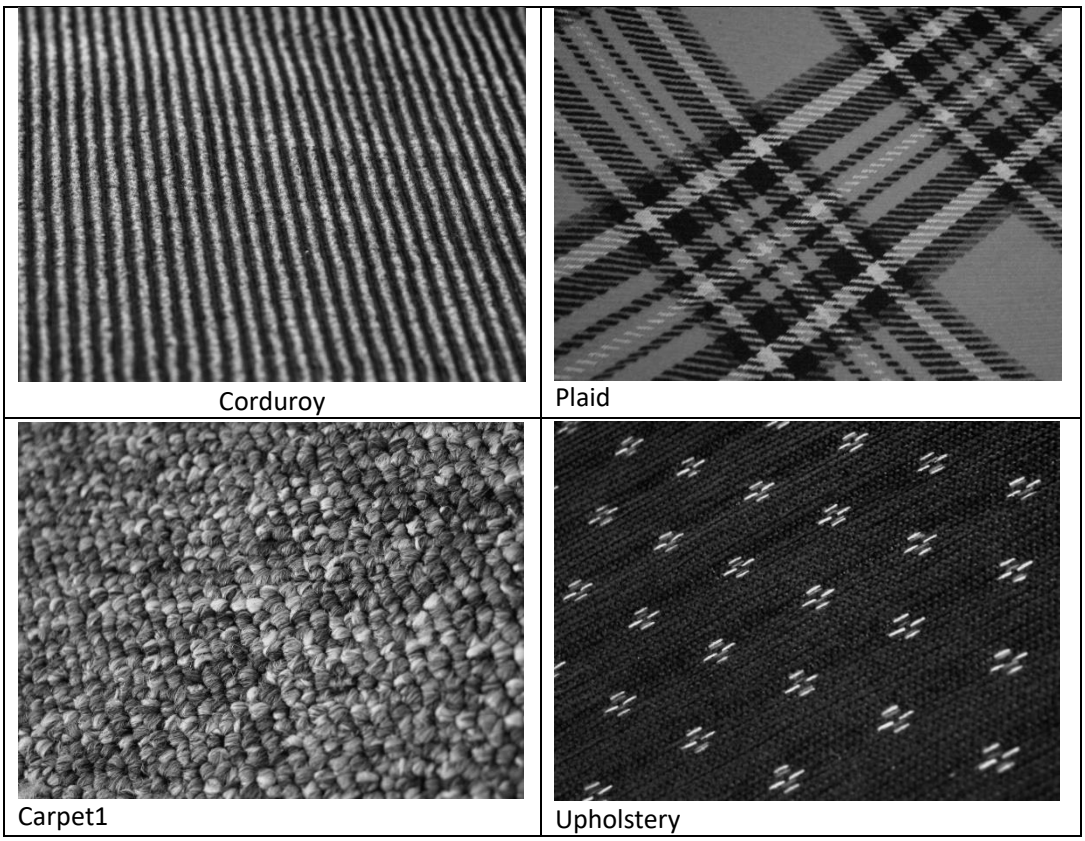
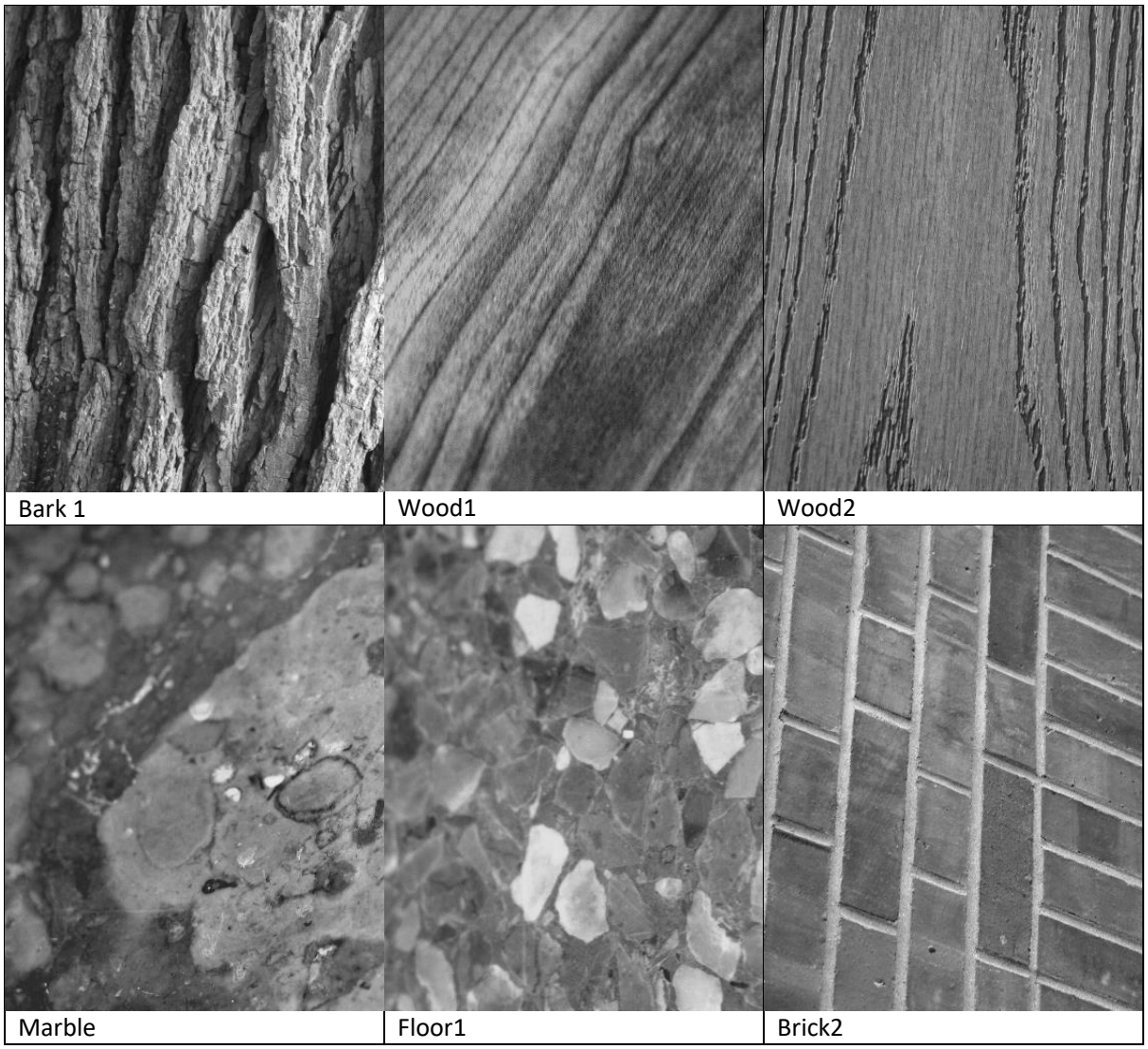


Figure 4.3 Texture classes with negligible effect of increase in N

On the other hand an increase in N has a large effect on the bark1, wood1, wood2, marble, floor1, brick2, carpet2, wallpaper and knit texture classes. These images have different patterns at different parts of the image. Also, there are many illumination changes within different parts of the image (Refer to Figure 4.4). It is observed that these textures have irregular patterns throughout the image. Thus Pixel N-grams with smaller value of N are unable to model these irregular textures; whereas, longer sequences can model the irregular patterns more appropriately providing high classification performance.



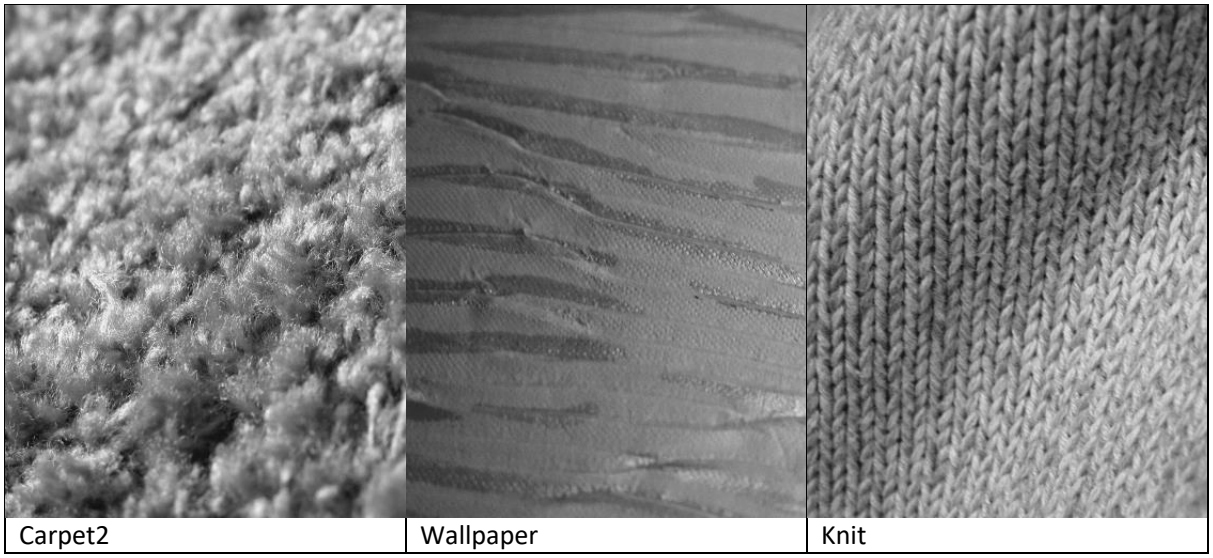


Figure 4.4 Texture classes with huge effect of increase in N

In order to understand the effect of N on regular and irregular textures let us consider a sample image of regular (consistently repeating) pattern and sample image of irregular pattern as shown in the Figure 4.5. These figures have only two grey levels black and white. Black is

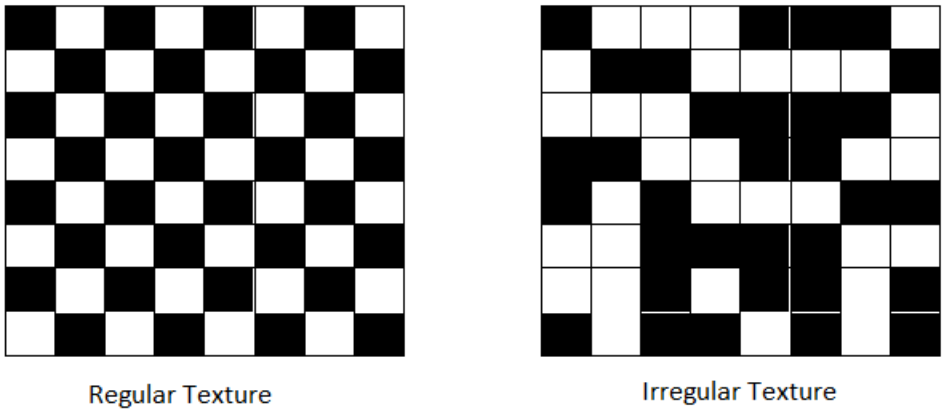


Figure 4.5 Example of regular and irregular texture pattern

considered 0 and white is 1. Figure 4.6 shows the 1-gram, 2-gram and 3-gram features for these two images. Consider each square as one pixel, thus the images are 8 pixel × 8 pixel in resolution. The number of black pixels are equal to the number of white pixels (32) in both the images for ease of comparison of N-gram features.

1-Gram	Count
0	32
1	32

2-Gram	Count
00	0
01	28
10	28
11	0

3-Gram	Count
000	0
001	0
010	24
011	0
100	0
101	24
110	0
111	0

Regular Texture

1-Gram	Count
0	32
1	32

2-Gram	Count
00	14
01	14
10	14
11	14

3-Gram	Count
000	5
001	8
010	6
011	6
100	8
101	2
110	6
111	5

Irregular Texture

Figure 4.6 N-gram features for regular and irregular texture patterns

It can be observed that for the regular texture image (checker board), there are two significant features for 1-gram, 2-gram as well as 3-gram representation. Thus it can be seen that increase in N does not provide any additional information for the regular texture images. Therefore, there is no significant difference between the classification accuracies (Fscore) obtained using 1-gram, 2-gram and 3-gram features for regular textures. On the other hand for the irregular texture image there are two 1-gram features, four 2-gram features and eight 3-gram features. Thus, as N is increased the N-gram features describe the image more accurately hence providing more accurate classification (High Fscore). This argument supports the experimental results for texture classification using various values of N described above.

4.1.2 Comparison with existing techniques (Classification performance)

In this experiment, the N-gram classification performance for texture image classification is compared with existing techniques including Intensity histogram, Haralick's features and BoVW features.

Firstly, Intensity histogram features were computed by counting the number of occurrence of particular grey level pixels in an image. These features were normalized using min-max normalisation and fed to SVM classifier.

Haralick's features (Haralick et al., 1973) based on co-occurrence matrix have been quite successful for texture classification. The four most effective Haralick's features useful for texture classification (contrast, correlation, energy and homogeneity) were computed using Matlab functions[8]. These features were then provided as input to a SVM classifier after min-max normalization.

The experiments on UIUC dataset for texture classification using BoVW approach have been conducted by Lazebnik (Lazebnik et al., 2005). The classification results are directly taken from this work for comparison purposes. The main parameter considered for comparison was Fscore (Please refer to Table 4.3). The generalization was estimated using 10 fold cross validation resampling.

Table 4.3 Texture image classification comparison

	Fscore			
Class	**Histogram**	**Haralick**	**BoVW**	**4-gram**
bark1	0.345	0.0	0.8972	0.789
bark2	0.633	0.155	0.8077	0.867
bark3	0.765	0.111	0.7455	0.935
wood1	0.114	0.000	0.9868	0.824
wood2	0.393	0.290	0.8983	0.921
wood3	0.581	0.000	0.9690	0.815
Water	0.587	0.037	0.9980	0.974
Granite	0.455	0.040	0.8352	0.828
Marble	0.267	0.352	0.7515	0.776
floor1	0.103	0.000	0.8500	0.918
floor2	0.637	0.170	0.8920	0.935
Pebbles	0.467	0.184	0.7947	0.961
Wall	0.787	0.220	0.9288	0.950
brick1	0.552	0.143	0.8307	0.895
brick2	0.049	0.000	0.8987	0.949

[8] Matlab 7.9.0 (R2009B)

	Fscore			
Class	**Histogram**	**Haralick**	**BoVW**	**4-gram**
glass1	0.529	0.000	0.7610	0.911
glass2	0.551	0.194	1.0000	0.964
carpet1	0.747	0.525	0.9660	0.976
carpet2	0.386	0.083	0.7270	0.911
Upholstery	0.964	0.988	0.9908	1.000
Wallpaper	0.175	0.143	0.7445	0.747
Fur	0.468	0.000	0.9453	0.789
Knit	0.161	0.000	0.8898	0.779
Corduroy	0.964	0.164	0.9937	0.988
Plaid	0.933	0.219	0.9330	0.987
Overall acc.	53.7%	21.9%	84.4%	**89.5%**

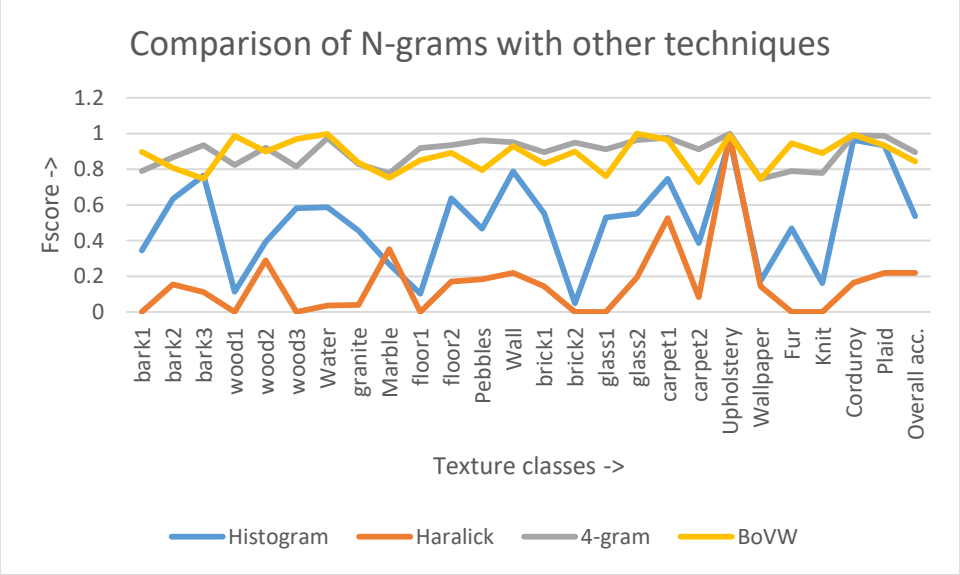


Figure 4.7 Comparison of Pixel N-grams with other techniques for texture classification (Fscore)

Figure 4.7 shows graph of variation in Fscore for various classes in the texture dataset using different techniques such as histogram, Haralick's features and BoVW. It can be observed that the Fscore with 4-grams was greater than the histogram and Haralick features for all the classes. Further, the Fscore using 4-grams was found to be better than BoVW approach for most of the classes except bark1, wood1, wood3, glass2, fur and knit. The overall classification accuracy using Pixel N-grams with 4-gram representation was 89.5% as compared to 84.4% using BoVW. One of the reasons that the Pixel N-grams performs better than BoVW is that the

information at every pixel level is important for texture classification which might be lost during the quantization/vocabulary construction for BoVW approach.

Table 4.4 notes the Precision and Recall values for classification using Histogram, Haralick and 4-gram features. BoVW results were taken from the work of (Lazebnik et al., 2005) which has only recorded the Fscore. The precision and recall values are not available in this work and hence they are not included in the Table 4.4 and Figure 4.9.

Table 4.4 Precision and Recall for texture dataset

Classes	**Precision**			**Recall**		
	Histogram	**Haralick**	**4-gram**	**Histogram**	**Haralick**	**4-gram**
bark1	0.556	0	0.833	0.25	0	0.75
bark2	0.641	0.118	0.837	0.625	0.225	0.9
bark3	0.929	0.1	0.973	0.65	0.125	0.9
wood1	0.133	0	0.778	0.1	0	0.875
wood2	0.571	0.345	0.972	0.3	0.25	0.875
wood3	0.543	0	0.805	0.625	0	0.825
water	0.519	0.071	1	0.675	0.025	0.95
granite	0.326	0.1	0.766	0.75	0.025	0.9
marble	0.4	0.232	0.655	0.2	0.725	0.95
floor1	0.167	0	0.867	0.075	0	0.975
floor2	0.569	0.119	0.973	0.725	0.3	0.9
pebbles	0.373	0.129	1	0.625	0.325	0.925
wall	0.685	0.214	0.95	0.925	0.225	0.95
brick1	0.511	0.121	0.944	0.6	0.175	0.85
brick2	1	0	0.974	0.025	0	0.925
glass1	0.643	0	0.923	0.45	0	0.9
glass2	0.388	0.126	0.93	0.95	0.425	1
carpet1	0.627	0.525	0.952	0.925	0.525	1
carpet2	0.647	0.094	0.923	0.275	0.075	0.9
upholstery	0.93	0.976	1	1	1	1
wallpaper	0.294	0.167	0.8	0.125	0.125	0.7
fur	0.407	0	0.903	0.55	0	0.7
knit	0.227	0	0.811	0.125	0	0.75
corduroy	0.93	0.238	0.976	1	0.125	1
plaid	1	0.516	1	0.875	0.8	0.975
Overall acc.	0.019	0.168	0.902	0.561	0.168	0.895

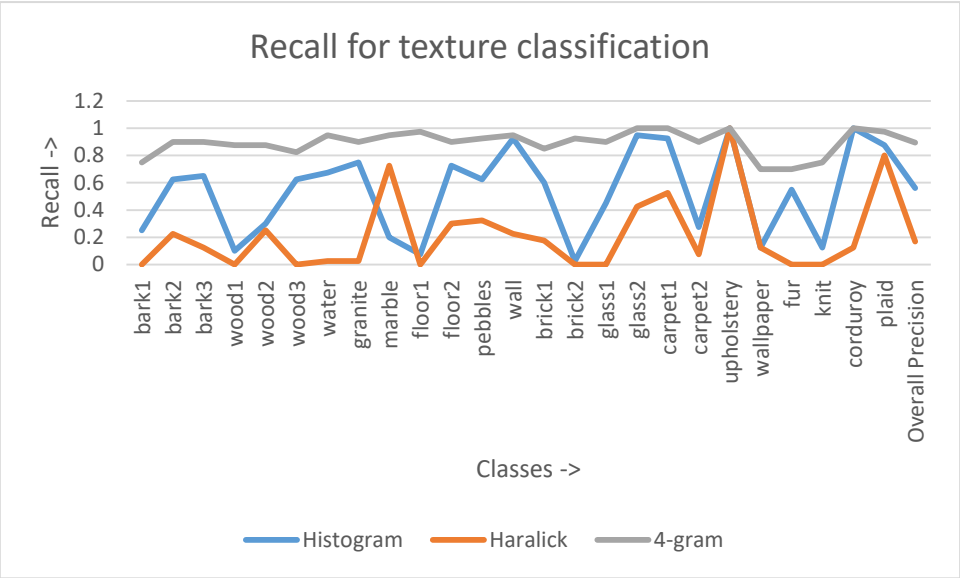


Figure 4.8 Comparison of Pixel N-grams with different techniques for texture classification (Recall)

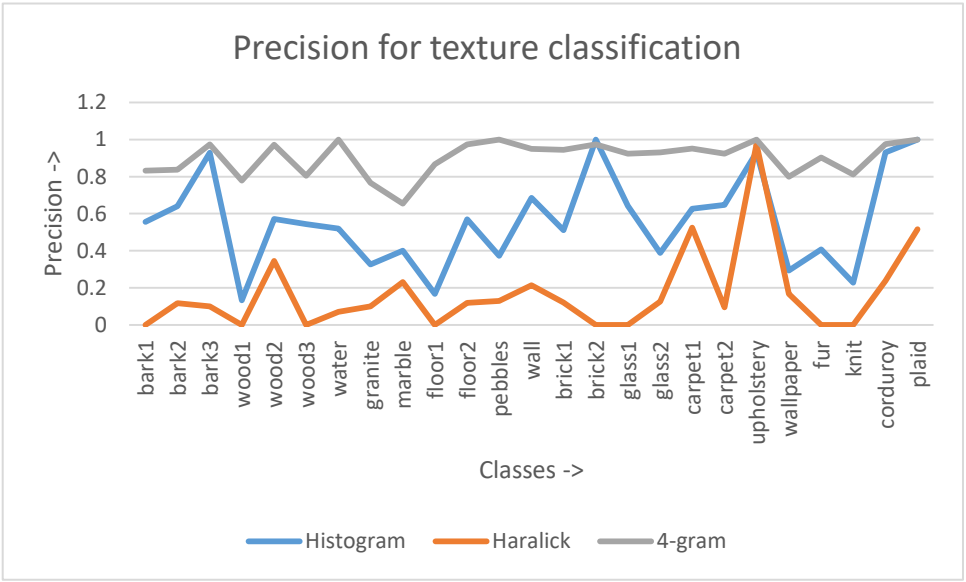


Figure 4.9 Comparison of Pixel N-grams with different techniques for texture classification (Precision)

Figure 4.9 and Figure 4.8 shows the precision and recall curves for different classes using histogram, Haralick's and 4-gram features. It can be observed that the precision as well as recall curve is steady for all the classes using 4-gram features, reinforcing that the 4-gram performance is independent of the classes. It also shows that the Pixel N-gram features give better classification performance with respect to precision and recall.

A series of two tailed paired t-tests were conducted in order to check whether the classification results using 4-gram features are significantly better than other techniques mentioned above in Table 4.3. The null hypothesis in each case is that the mean of the Fscores using 4-gram features is the same as the mean of the Fscores using other features. The level of significance was set

at $\alpha = 0.05$. If the p-value is less than the level of significance then the null hypothesis that there is no difference between groups is rejected. The p-values obtained for the paired t-tests using various features were noted in the Table 4.5.

Table 4.5 T-test results for comparing texture classification

Feature sets used	**p-value**
4-gram vs Histogram	0.00
4-gram vs Haralick	0.00
4-gram vs BoVW	0.48

The p-values for the Intensity Histogram and Haralick features were less than the level of significance hence we can conclude that the classification results using 4-gram features are significantly better than those using Histogram and Haralick features. The p-value obtained by paired t-test of 4-gram results and BoVW approach indicates that the 4-gram features give comparable performance with BoVW approach.

4.1.3 Comparison with existing techniques (Computational complexity)

The main advantage of Pixel N-grams approach is the ease of use and computational cost. Calculating Pixel N-gram features only involve counting the occurrences of the various N-grams present in an image. For 4-gram computation with images reduced using 8 grey level bins, the dimension of the feature vector becomes $8^4 = 4096$. Whereas, the computation of BoVW approach (Lazebnik et al., 2005) involves many steps such as computing the salient keypoints in an image, building a visual vocabulary, and finding out the closest visual word from the codebook. The number of keypoints could be as few as 100 or as high as 1000 depending on the complexity of an image. Each keypoint then needs to be described with the help of Scale Invariant Feature Transform (SIFT) descriptors. To calculate the SIFT descriptor (Lazebnik et al., 2005; Lowe, 2004), first a set of orientation histograms was created on typically a 4x4 pixel neighborhoods with 8 bins each. Histograms are computed from magnitude and orientation values of samples in a 16 x 16 region around the keypoint. The magnitudes need to be further weighted by a Gaussian function with σ equal to one half the width of the descriptor window. The descriptor then becomes a vector of all the values of these histograms. Thus 4 x 4 = 16 histograms each with 8 bins produce a feature vector of 128 elements. In a complex image, thus the feature vector dimension increases to 128 × 1000 = 128,000. The SIFT features are then clustered normally using k-means algorithm. The centroids of the clusters are considered visual words and are used to build a visual words codebook or dictionary/vocabulary. The image is then represented using the frequency of occurrence of the visual words present in it.

The steps for computation of BoVW features and Pixel N-gram features clearly indicate that the BoVW representation using a keypoint approach is computationally very expensive as compared to Pixel N-gram representation.

Further, the Haralick features are second order statistical features and hence require more time to compute. Considering that the Pixel N-grams give significantly better performance than the Haralick's features and comparable performance to BoVW with very little computational cost and ease of use, this model is a promising approach for texture image classification applications.

4.2 Experiments on Shapes Dataset

Along with texture, shape is another important characteristics of breast lesions (Wei et al., 2011). Further, lesions can be of different sizes and located at different locations in a mammogram. Moreover, mammographic images taken on different devices have varying resolutions. Classification of mammographic lesions irrespective of image resolutions could be very useful for automated breast cancer detection and diagnosis and help the clinicians.

Motivated by these observations this section reports on the application of the Pixel N-grams model to the detection of predefined shapes in an image, regardless of the size or location of the shapes. Experiments were conducted on a database of binary shape images to examine the extent to which classification using Pixel N-grams may be independent of size and location of a shape in an image. Further, verification of the resolution independency of these features is also important. The increase in N has a significant effect on classification accuracy, therefore experiments to find out the optimum value of N for the shapes database were also conducted. The details of the experiments are detailed in the subsections.

Results of the experiments on shapes dataset were published in (P. Kulkarni et al., 2016).

4.2.1 Finding optimum value of N for shapes

A change in value of N may have a big effect on the classification performance. In order to find out the optimum value of N, the classification of basic shapes using different values of N was observed. The database consisting of 240 binary images of three basic shapes (80 images each) was used for this experiment. The classification performance was measured by considering Precision, Recall and Fscore criteria. The results were noted in the Table 4.6.

Table 4.6 Optimum value of N for shapes dataset

	Triangle			Circle			Square		
	Precision (%)	Recall (%)	Fscore (%)	Precision (%)	Recall (%)	Fscore (%)	Precision (%)	Recall (%)	Fscore (%)
Onegram	79.8	98.8	88.3	93.3	87.5	90.3	93.9	77.5	84.9
Twogram	97.5	96.3	96.9	100	95.0	97.4	94.1	100	97.0
Threegram	100	100	100	100	100	100	100	100	100

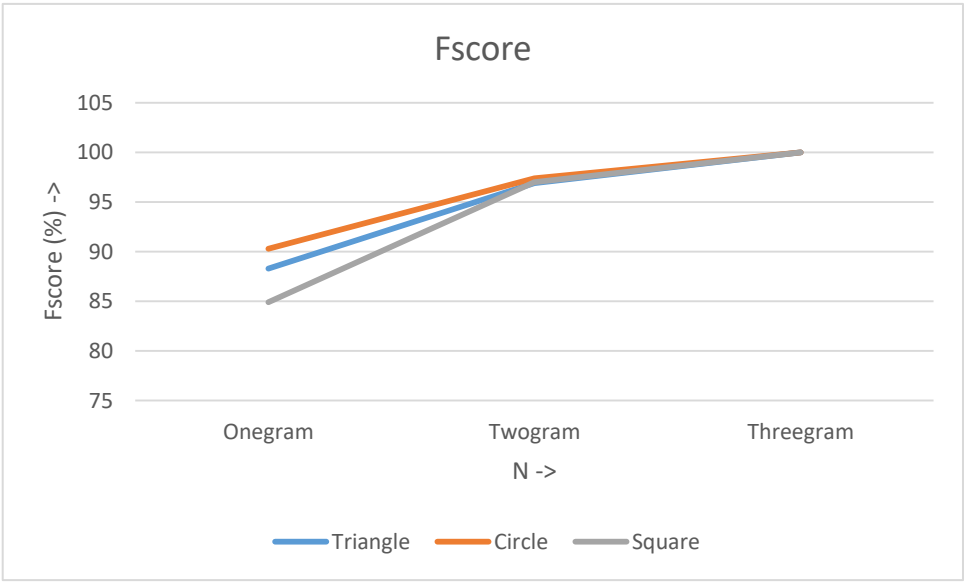


Figure 4.10 Effect of N on Fscore for shapes dataset

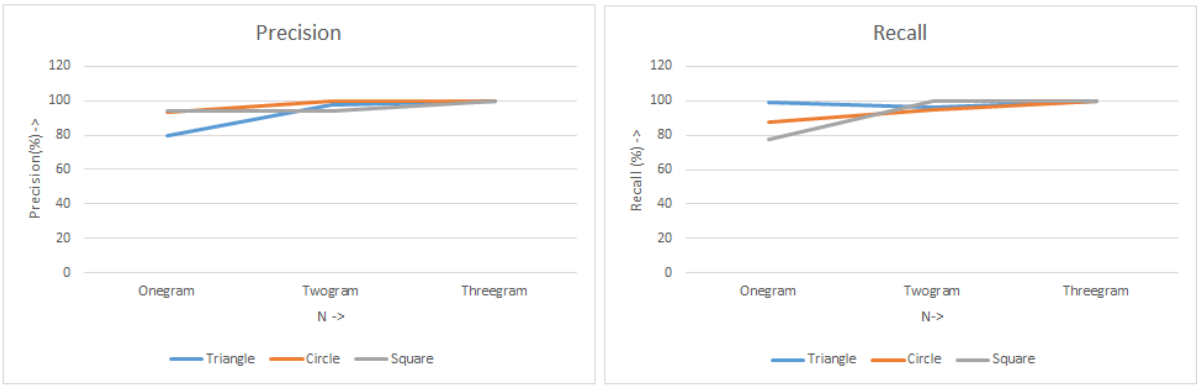


Figure 4.11 Effect of varying N on precision and recall for shapes database

Figure 4.11 shows the effect of increasing N on the Fscore and Figure 4.10 shows the effect of increasing N on the precision and recall for various shapes. It can be observed that the precision, recall as well as Fscore increased as N is increased from 1 to 3. At N=3, perfect classification with 100% precision, 100% recall and 100% Fscore was achieved. Thus, further increase in N were not explored. Therefore, the optimum value of N for this binary shapes dataset was set at 3. Hence all the shape classification experiments mentioned henceforth were conducted using N=3.

4.2.2 Size invariance

For this experiment 60 images from the above mentioned shapes database were used. Three shapes (circle, triangle and square) of 20 different sizes each at centre location were considered. These images were of size 512 × 512 pixels. Some example images are shown in the Figure 4.12.

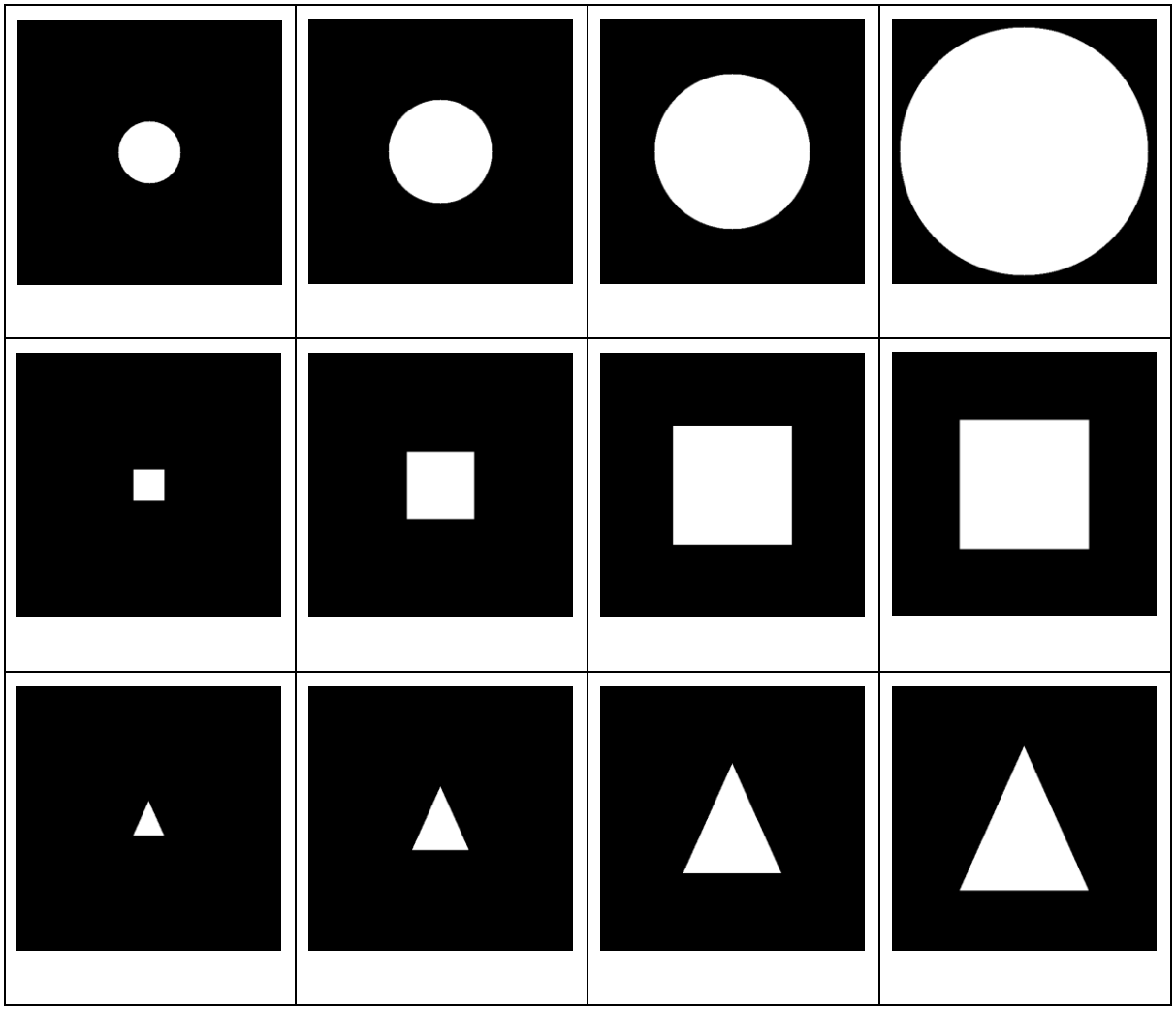


Figure 4.12 Example shape images of different size

It is surmised that the basic shapes could be described by providing details in horizontal and vertical directions. Therefore, 3-gram features in horizontal + vertical direction were computed. The images are binary and contain only two grey levels (black is considered 0 and white is considered 1), we get the following N-grams for shape images: 000, 001, 010, 100, 110, 011, 111. The N-gram counts were normalized using min-max normalization and provided as input to the MLP classifier. Then intensity histogram features were calculated and normalized. These normalised histogram features were fed to the MLP classifier for classification into 3 classes (circle, triangle and square).

Further, for comparison purpose four main Haralick's features (Haralick et al., 1973) (co-occurrence matrix based) contrast, correlation, energy and homogeneity were calculated. These features were normalized and used for classification using MLP classifier. Due to the small dataset size leave one out validation was used to estimate the generalisation. Results were noted in Table 4.7.

Table 4.7 Shape classification with different sizes

Features	Triangle (%)			Circle (%)			Square (%)			Overall		
	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall
Histogram	78.3	66.7	94.7	66.7	100	50	61.9	59.1	65	87.9	89	87
Haralick	50	50	50	100	100	100	50	50	50	66.67	50	67
Pixel N-grams	100	100	100	100	100	100	100	100	100	100	100	100

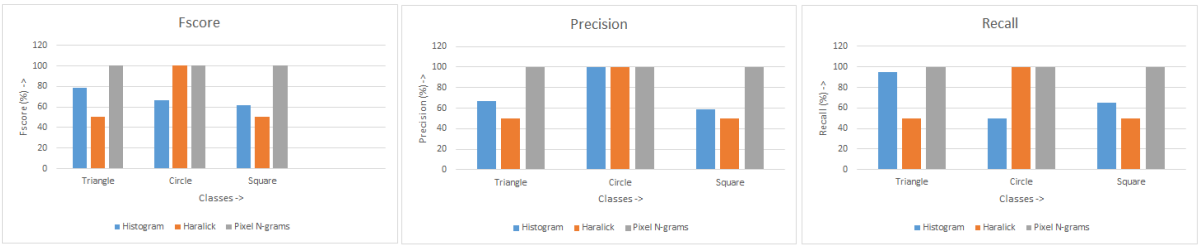


Figure 4.13 Shape classification (size invariance)

The shape classification results using different sizes of triangles, circles and squares using intensity histogram, Haralick's features and Pixel N-gram features are graphically shown in Figure 4.13. It can be seen from these graphs that the Pixel N-gram features (horizontal + vertical direction) using the MLP classifier were able to perfectly classify the shape images of different sizes (Fscore of 100% with 100% precision and recall). The results were compared with existing techniques such as intensity histogram and Haralick's features using Fscore, Precision and Recall (Refer to Table 4.7).

It was observed that for the circle class, the performance of Pixel N-grams was comparable with that of the Haralick's features. Two tailed paired T-tests run on the classification results using histogram and Pixel N-gram features resulted in p value of 0.0008. Similarly, the two tailed paired T-test run on classification results using Haralick's features and Pixel N-gram features resulted in p value of 0.00395. These p-values are less than the level of significance $\alpha = 0.05$. This suggests that the classification using Pixel N-gram features work significantly better than the classification using Intensity histogram and Haralick's features for shapes of different sizes. Thus it can be concluded that the shape classification using Pixel N-gram features is relatively size invariant for three simple shapes.

4.2.3 Location invariance

The objective of this experiment is to see how accurately N-gram features can classify shapes located at different sites in an image. For this experiment, various images were created by changing the shape locations. This was achieved by varying the shape's centre co-ordinates along the X axis, along the Y- axis and then along the diagonals. The dataset thus consisted of 80 circles, 80 triangles and 80 squares at different locations. All images were of size 512 × 512 pixels. Some example images are shown in the Figure 4.14.

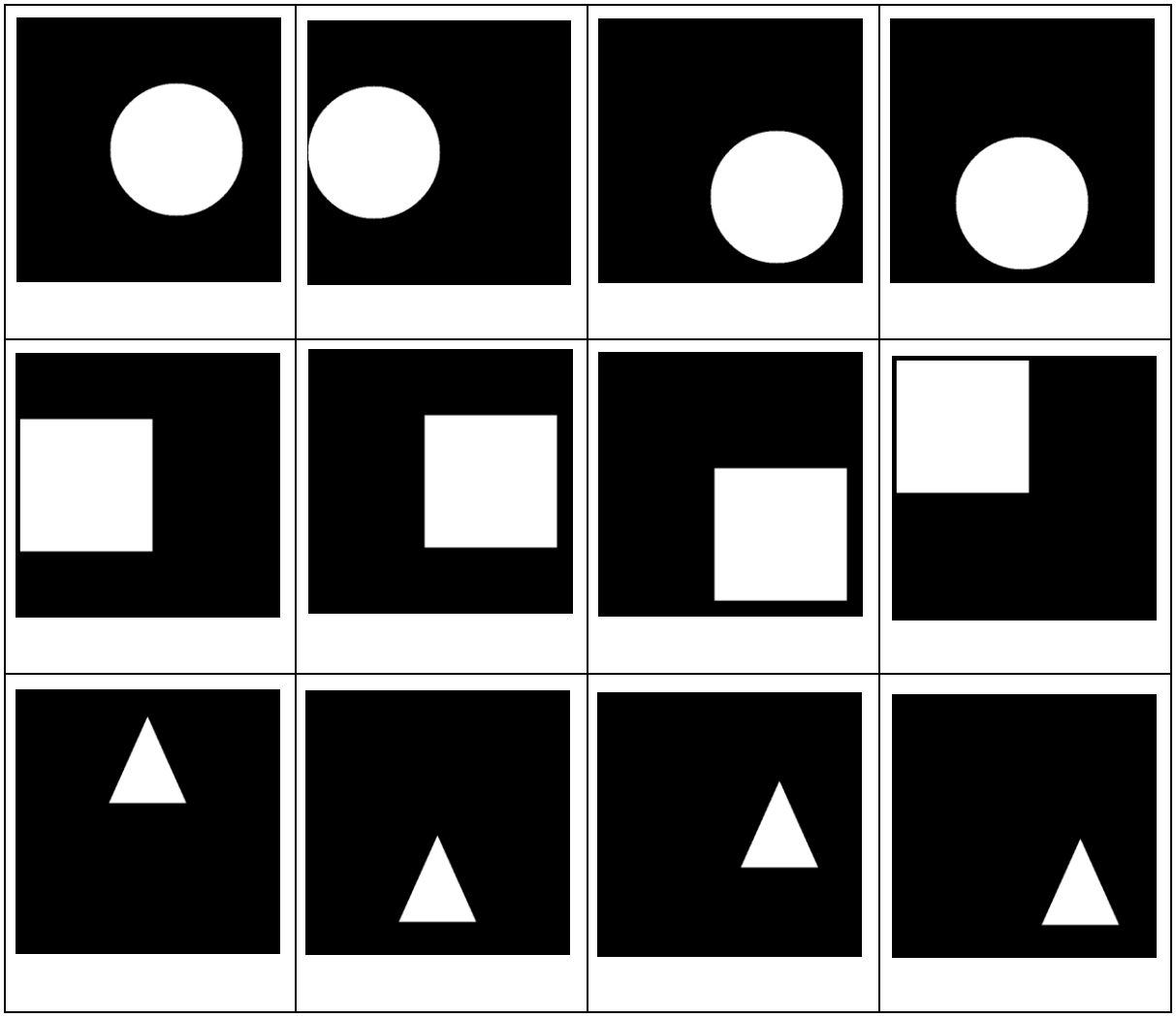


Figure 4.14 Sample shapes at different locations

3-gram features in horizontal as well as vertical directions were computed and given as input to the MLP classifier. Leave one out validation was carried out to verify the classification accuracy due to smaller data size. The results were noted in Table 4.8. Fscore with Histogram features was noted as 88.9% for triangle class, 88.9% for circle class, 83.0% for square class with an overall Fscore of 87.1%. The Fscore noted for Haralick's features was 66.7% for

triangle class, 100% for circle class and 0% for square class with an overall Fscore of 55.6%. Whereas, the Pixel N-grams provide Fscore of 100% for all the classes.

Table 4.8 Shape classification results (shapes at different locations in image)

Features	Traiangle (%)			Circle (%)			Square (%)			Overall		
	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall
Histogram	88.9	80	100	88.9	93.2	85.0	83	91	76.3	87.1	88.1	87.1
Haralick	66.7	50	100	100	100	100	0	0	0	55.6	66.7	66.7
Pixel N-grams	100	100	100	100	100	100	100	100	100	100	100	100

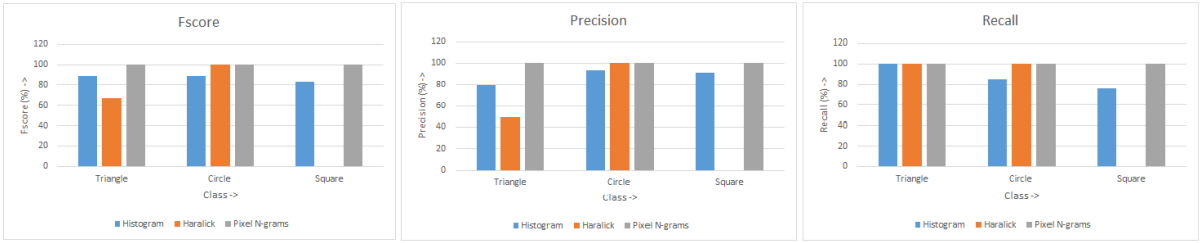


Figure 4.15 Shape classification with different locations

Figure 4.15 shows the classification performance graphically for various features (Histogram, Haralick, Pixel N-grams) when the shapes are present at various locations in an image. It is evident from Table 4.8 that Haralick's features are able to classify circles at different locations accurately however fail to classify the squares at different locations. The Pixel N-grams give steady and better performance than histogram and Haralick's features for all the shape classes. Thus it can be concluded that the Pixel N-gram features can classify three simple shapes irrespective of the location of the shape in an image.

4.2.4 Resolution invariance

As explained earlier, the mammograms generated using different x-ray machines may yield images of different resolutions depending on the settings and specifications of the machinery used. If the repository contains mammographic images of different resolutions, automatic detection systems should be able to classify the images accurately irrespective of the resolution of the image. This experiment has been designed to assess the extent to which the Pixel N-gram features can be used to classify shape images of varying resolutions.

For simplicity, images of 3 basic shapes (triangle, circle and square) of 10 different resolutions (512 × 512 , 1024 × 1024, 1536 × 1536, 2048 × 2048, 2560 × 2560, 3072 × 3072, 3584 × 3584, 4096 × 4096, 4608 × 4608, 5120 × 5120) were generated for this experiment. It was determined with the experiment described in Section 4.2.1t hat N=3 is the optimum value of N for the shapes dataset. Therefore, 3-gram features in the horizontal and vertical direction were computed for every image in the dataset. These features were then normalised using min-max

normalisation and fed to the MLP classifier. Again, leave one out validation was used to estimate the generalisation due to smaller data size. Classification accuracy (%) was compared with histogram and Haralick features (See Table 4.9).

The BoVW approach using SIFT features has proven to be size, location invariant (Setitra & Larabi, 2015) and resolution/scale invariant (Lowe, 2004) and hence the performance of BoVW approach has not been examined for size, shape and resolution invariance.

Table 4.9 Shape classification results (different resolution images)

Features	Triangle (%)			Circle (%)			Square (%)			Overall		
	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall	Fscore	Prec	Recall
Histogram	84.2	88.9	80.0	76.2	72.7	80	73.7	77.8	70	80	80.2	80
Haralick	0	0	0	100	100	100	9.5	9.1	10	36.5	36.4	36.7
Pixel N-grams	90	90	90	100	90.9	100	90	90	90	90	89.9	90

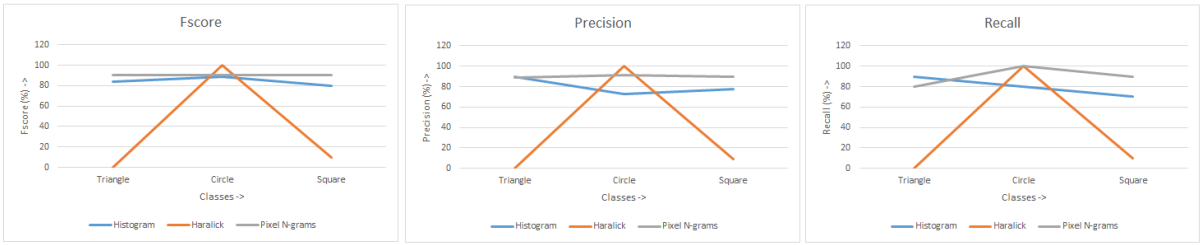


Figure 4.16 Shape classification performance with different resolution images

It is observed that Pixel N-gram features provide classification accuracy of 84.2% for triangle class, 95.2% for circle class, 90% for square class with an overall accuracy of 90% which is higher than that using histogram and Haralick's features.

Figure 4.16 shows the variation in the classification performance (Fscore, precision and recall) for all the classes using various feature extraction techniques (Intensity histogram, Haralick's features and Pixel N-gram features) for images of different resolution. It was observed that the Pixel N-grams are able to classify shape images irrespective of the resolution of the image reasonably well. The classification performance using Pixel N-grams for circle class was comparable with that of Haralick's features. The overall classification accuracy for shape classification with different resolutions was found to be better than the Intensity histogram and Haralick's features. Further, the classification accuracy is stable for all the shapes using Pixel N-gram features. The two tailed t-tests conducted obtain p value of 0.000453 for histogram and Pixel N-gram results whereas p value of 0.005461 for Haralick and Pixel N-gram features indicating that the performance of Pixel N-grams is significantly better than these two feature extraction techniques for shape classification with different image resolutions.

4.3 Chapter Summary

Texture and shape are the two important characteristics useful for mammographic lesion classification. In this chapter use of Pixel N-gram model for texture classification and shape classification has been explored.

The experiments on UIUC texture dataset suggest that as N is increased the classification performance for irregular texture images is improved whereas, the N has negligible effect on the classification performance of the regular textures. The 4-gram features were observed to give the best performance for texture classification from UIUC texture database. Thus the optimum value of N for UIUC texture dataset was noted as N=4. It was also observed that the Pixel N-grams provide significantly better classification performance as compared to Intensity histogram and Haralick's features for texture classification. Further, the classification performance of Pixel N-grams was found to be comparable with that of the BoVW approach, with an added advantage of simplicity and low computation cost.

Apart from the biomedical image classification for disease detection and diagnosis, there are several other applications of texture image classification such as ground classification and segmentation of satellite or aerial imagery, segmentation of textured regions in document analysis, and content-based access to image databases. However, despite many potential areas of application for texture analysis there is only a limited number of successful applications. A main problem is that textures in the real world are often non uniform, due to changes in orientation or scale. Additionally, most of the existing texture extraction methods are computationally complex. The Pixel N-gram technique is found to be computationally less expensive than existing texture feature extraction techniques suggesting further research in this direction may be promising.

Experiments on the binary shapes dataset demonstrate that the Pixel N-gram features were able to distinguish among various basic shapes accurately (100% accuracy, 100% precision and 100% recall). For the shapes dataset, N=3 provided the best classification performance and therefore, N=3 was noted as the optimum value of N for binary shapes dataset. Moreover, the classification using Pixel N-grams was found to be independent of the size and location of the shapes in an image. Further, the shape classification using Pixel N-grams was noted to be resolution independent. Also, the classification performance of Pixel N-grams for shape classification was significantly better than the Intensity histogram and Haralick features.

Shape classification is very useful for biomedical image classification as well as other object detection applications. Pixel N-gram features provided promising results for the basic shapes dataset opening up door for many shape classification applications. However, further investigation with complex shape classification needs to be performed in order to investigate use of Pixel N-grams for different shape classification applications.

5 Experimental Results and Analysis (Mammographic Images)

This chapter details experiments carried out to test the efficacy of the novel Pixel N-gram features. The main purpose of the experiments is to analyse the use of Pixel N-grams for mammographic lesion classification.

Two types of classification of lesions were used: normal/abnormal classification and circumscribed/speculation/normal classification. Classification performance for all experiments was measured using Fscore, Sensitivity and Specificity. Grey scale reduction was performed in order to reduce the computational cost. Piecewise constant approximation of the Pixel N-grams with 256 grey levels was also tried for solving the problem of optimum grey level selection. Similarly, various classifier performances were analysed in order to determine the performance of different classifiers for mammographic lesion classification using Pixel N-gram features. Further, wrapper approach was used to select the best features for improving the classification performance and reducing feature dimensions. Finally, the computational complexity of Pixel N-grams was analysed and compared with existing techniques.

For the details of the datasets and methodology for classification of mammograms please refer to Chapter 3 (Research Methodology). Two mammographic datasets miniMIAS and Lakeimaging were used for the experiments. The classification results on the miniMIAS dataset were published in two international conferences (P. Kulkarni et al., Feb 2014; P. Kulkarni, Stranieri, A., Kulkarni, S., Ugon, J., & Mittal, M., 2015) and a peer reviewed journal (P. Kulkarni et al., Mar 2014). The results on LakeImaging dataset were published in an international conference (P. Kulkarni, Stranieri, A., Ugon, J., Kulkarni, S & Mittal, M., April 2017). The experiments carried out include:

1. Finding optimum value of N and Grey scale reduction
2. Effect of using different types of binning strategies
3. Comparison of different classifiers
4. Effect of choosing different normalisation techniques
5. Normal/Abnormal Classification
6. Circumscribed/Speculation/Normal Classification
7. Feature selection using wrapper approach
8. Piecewise constant approximation
9. Computational complexity comparison

5.1 Finding Optimum Value of Grey Scale Reduction

It is highly desirable to have low computational cost for mammographic classification algorithm in order to increase the efficiency of the radiologists. By reducing the images in grey scales the number of N-grams to be computed are reduced. Thus the computational cost can certainly be reduced. Sample grey scale reduced images of ROI's (circumscribed, speculated and normal) using different values of grey scales can be seen in Figure 5.1.

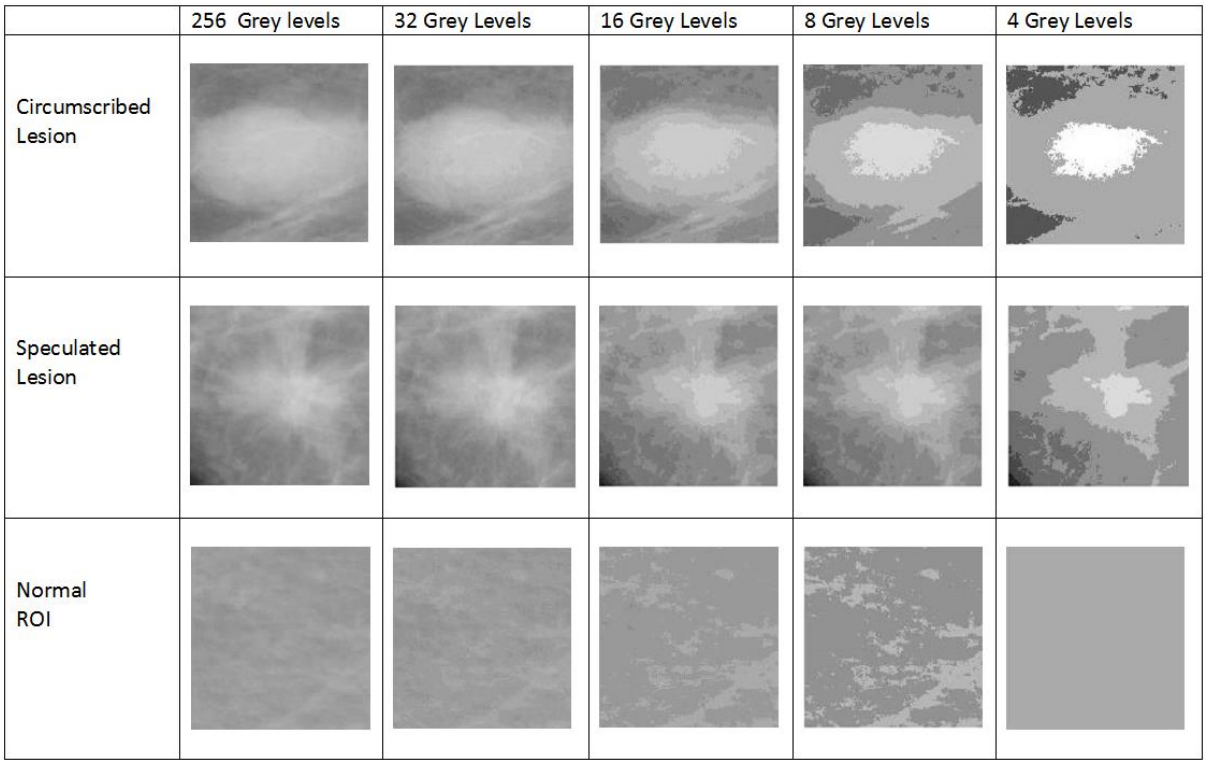


Figure 5.1 Grey scale reduced ROI's

Also, it is clear that the possible number of N-grams varies with the grey scale reduction. The maximum possible dimension of the feature vector formed using N-grams in a particular direction is given by equation 5.1, where Ng = number of grey levels and N= Number of adjacent pixels considered for computing N-grams.

$$Feature\ vector\ dimension = (N_g)^N \tag{5.1}$$

Figure 5.2 shows how the image is reduced in grey scale using 8 grey levels. Thus with the images reduced to 8 grey level bins, 8 one gram features, 64 two gram features, 512 three gram features and 4096 four gram features for every ROI can possibly be generated.

30	55	62	131	140	102	145
28	48	25	75	78	83	58
25	180	200	205	104	108	88
63	196	220	210	99	74	69

Original Image

1	2	2	5	5	4	5
1	2	1	3	3	3	2
1	6	7	7	4	4	3
3	7	8	8	4	3	3

Gray scale reduced image with 8 bins

Figure 5.2 Grey scale reduction using 8 grey levels

However, it has been observed that not all the possible N-grams are present in the given corpus which further reduces the dimension of the feature vector. Table 5.1 shows the possible number of N-grams and number of N-grams actually present in the miniMIAS corpus at a particular value of grey scale reduction for N=3.

Table 5.1 Possible and actual 3-grams with different grey scale reduction

Grey Scales	Number of possible 3-grams	Number of 3-grams present in corpus	Proportion of possible N-grams
256	16,777,216	4334	0.00258%
128	2,097,152	1268	0.0604%
64	262,144	403	0.1537%
32	32768	137	0.418%
16	4096	82	2.00%
8	512	35	6.84%
4	64	25	39.06%
2	8	8	100%

The observed number of N-grams with respect to number of grey levels fit on a polynomial relationship given by equation 5.2 and the proportion of possible N-grams that are observed can be modelled by equation 5.3. The graph of observed number of N-grams versus number of grey scales and that of proportion of observed N-grams versus number of grey scales is shown in the Figure 5.3.

$$N_{observed} = 0.0586G^2 + 2.2623G + 13.371 \quad (5.2)$$

$$N_{proportion} = 5.5066G^{-2.048} \quad (5.3)$$

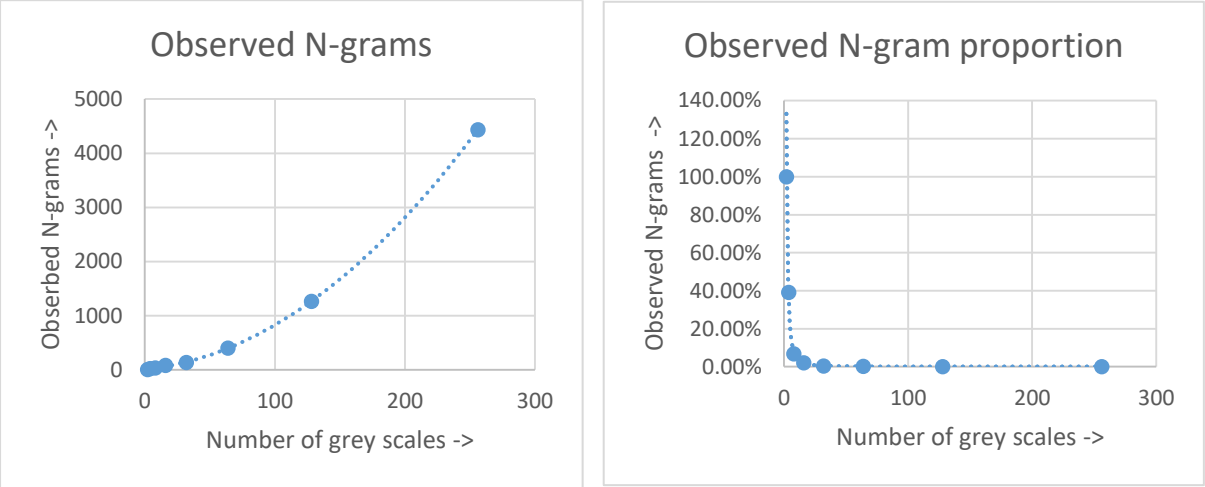


Figure 5.3 Trend in observed N-grams as a function of grey scale reduction

From these two graphs in Figure 5.3, it is clearly seen that the proportion of possible N-grams which are actually observed decreases with the increase in number of grey scales used. The decrease in observed N-grams is not linear as the possible number of N-grams increase cubically with increase in number of grey scales. It is arguable that a classifier trained with observed data generalises to classify all possible points if the observed data represents a larger proportion of the theoretically possible points. Thus, there is a trade-off between wanting to decrease grey levels so that proportion of observed/possible is higher and wanting to increase grey level so that the features are not too coarse and ineffective for distinguishing between classes. Since the observed N-grams are used for training the classifier for mammographic classification, the classifier would generalise better for grey scale reduction to two grey levels, as observed N-gram proportion is 100% in this case. However, grey scale reduction to two grey levels might incorporate quite a bit of information loss resulting in too coarse features. On the other hand the classifier using N-gram features without grey scale reduction (256 grey levels) might provide higher classification performance but would be less effective for generalization as it is trained using 0.00258% of all possible N-grams. Therefore, the best grey level to select is the level on the graph where the proportion of the possible N-grams which are actually observed begins to drop dramatically. Consequently, the grey scale reduction to 8 grey levels was chosen.

5.2 Finding Optimum value of N

As the value of N is increased, the Pixel N-gram representation of image becomes more and more complete. The complete representation of an image is obtained if the spatial relationship of every pixel with every other pixel is modelled while generating features. However, with

increase in N the dimensionality of the feature vector is increased producing the risk of overfitting which is normally referred to as 'Curse of dimensionality problem' (Bankman, 2008). Due to this the classification performance could be degraded. Additionally, as the longer sequences are hardly observed, with increase in N the feature vector becomes too specific to a particular image making it hard for the classifier to generalise well. Moreover, the computational cost is increased with the increase in N. Thus a balance has to be achieved between increasing N for achieving the complete image representation (and therefore better classification performance) and decreasing N in order to reduce the dimensionality of the feature vector (improve classifier generalisation, avoid increase in computational cost). It is therefore necessary to obtain an optimum value of N. For finding the optimum value of N, classification performance is analysed with change in value of N. Values of N considered for the experiment were 1, 2, 3, 4 and 5. With further increase in N the vocabulary size and feature vector size is increased, thus increasing the computational cost. Further, with increase in N the sequences are hard to find therefore resulting in features more specific to a particular image. This will make it harder for the classifier to generalize. Therefore, we stop at N=5. The images were grey scale reduced using 8 grey levels as this was found to be the optimum grey scale reduction value.

Table 5.2 Effect of varying N (miniMIAS dataset)

Performance criteria	1-gram (%)	2-gram (%)	3-gram (%)	4-gram (%)	5-gram (%)
Fscore	72.5	80.2	86.4	85.2	79.6
Sensitivity	73.0	79.9	84.6	82.2	80.2
Specificity	88.2	88.0	88.9	88.3	88.5
ROC area	80.3	83.4	85.8	84.1	81.6

Table 5.3 Effect of varying N (LakeImaging dataset)

Performance criteria	1-gram (%)	2-gram (%)	3-gram (%)	4-gram (%)	5-gram (%)
Fscore	75.3	82.7	83.9	79.0	78.8
Sensitivity	75.0	83.2	84.0	79.0	78.3
Specificity	91.4	89.3	91.1	89.8	90.3
ROC area	84.5	86.8	90.6	88.0	86.2

The experimental results to determine the optimum value of N for miniMIAS as well as LakeImaging dataset using the circumscribed/speculated/normal classification are given in the Table 5.2 and Table 5.3.

The graphical representation of effect of varying N on the classification performance for miniMIAS and LakeImaging dataset can be seen in Figure 5.4 and Figure 5.5 respectively. For comparison of effect of N on each of the different performance criteria (Fscore, sensitivity, specificity, Receiver Operating Characteristic curve area) on both the datasets please refer to graphs shown in Figure 5.6.

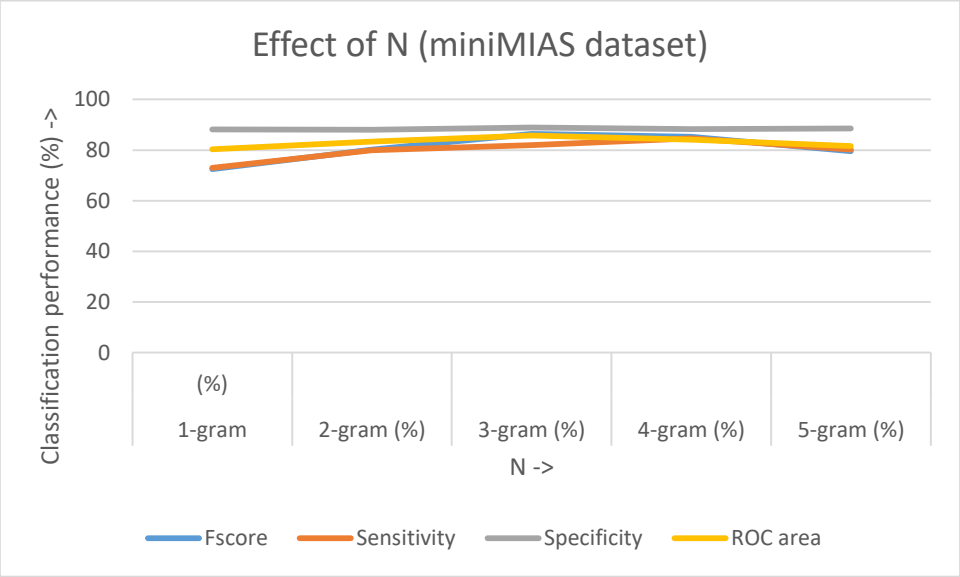


Figure 5.4 Effect of N on classification performance (miniMIAS)

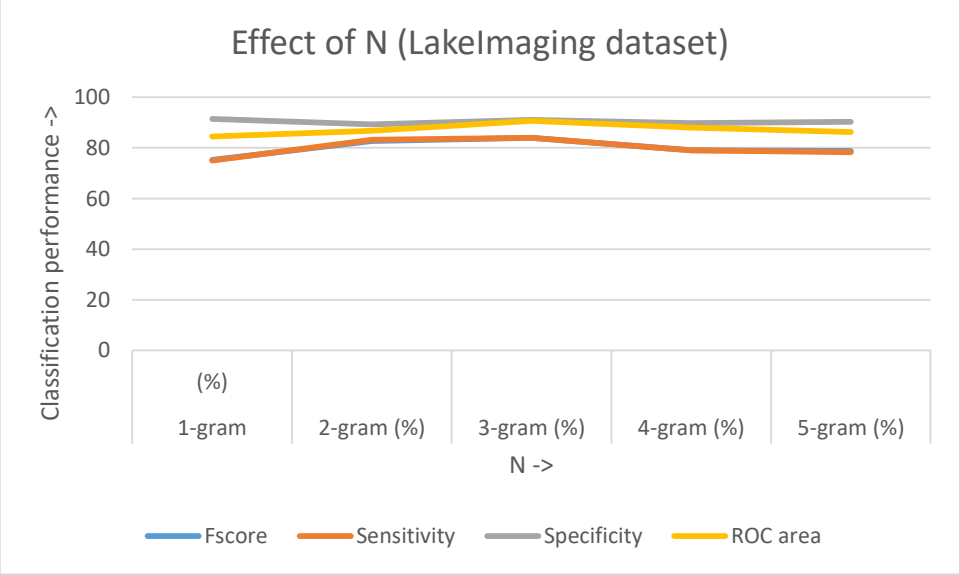


Figure 5.5 Effect of N on classification performance (LakeImaging)

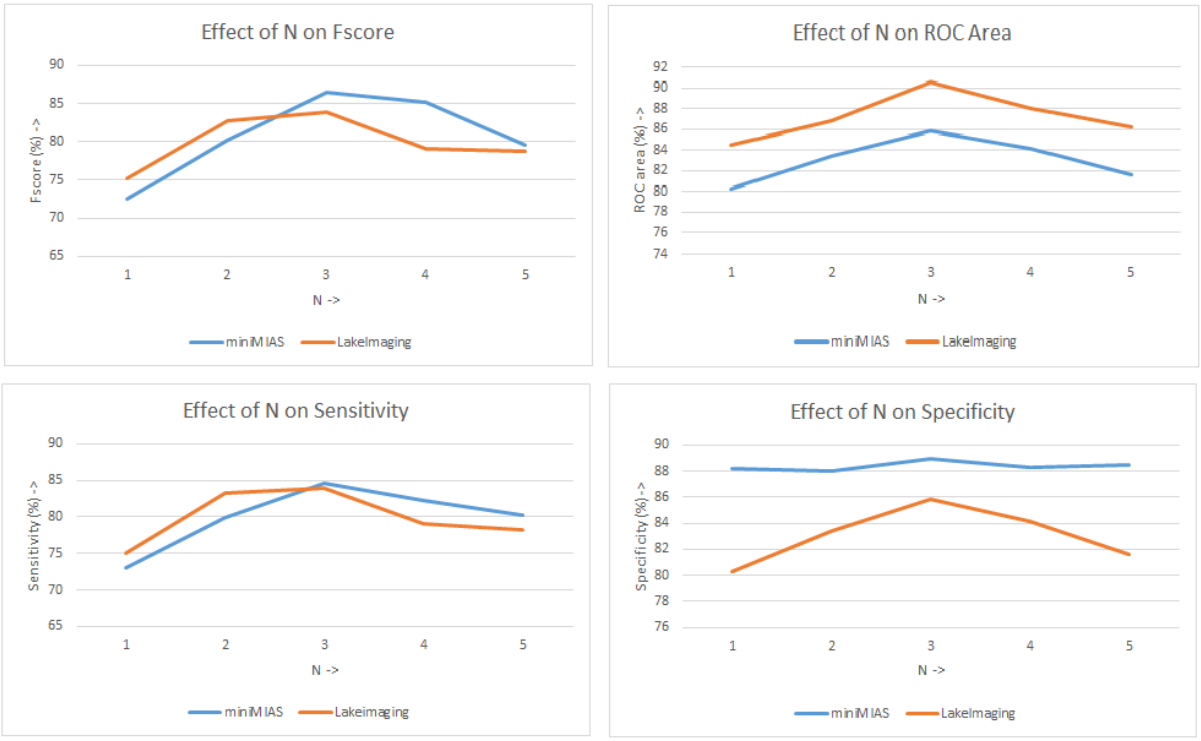


Figure 5.6 Effect of N on miniMIAS and LakeImaging dataset

From Table 5.2 and Table 5.3 it is evident that the classification accuracy increases when N is increased up to 3, however starts to fall as 'N' is further increased to 4 and 5 for both the datasets (miniMIAS and LakeImaging). Similar observation could be found in case of sensitivity. The sensitivity has higher value for N=3 as compared to that for N=1, 2, 4 and 5. Also, area under the receiver operating characteristic curve (ROC area) increases with increase in N upto N=3 and then starts decreasing. Similarly, specificity values are found to be higher at N=3 for LakeImaging dataset. However, the specificity values seem to have small variation with change in N for miniMIAS dataset. Thus the optimum value of N for miniMIAS as well as LakeImaging dataset is 3. Therefore, for all the further experiments value of N=3 and grey scale reduction G=8 was finalised as optimum values to use.

5.3 Effect of using Different types of Binning Strategies

Combining the different intensity levels into one intensity level for the purpose of computational cost reduction and noise reduction is known as binning. As discussed in the Chapter 3, mainly two types of binning strategies can be used for reduction of images in grey scales. One is 'Equal Size Binning' where there are same number of intensity levels in every bin. For example bin1 from 0 to 31, bin2 from 32 to 63 and so on. Another binning strategy is 'Equal Frequency Binning' where, the number of pixels in every bin are same but the start and the end grey scales of bins can vary from bin to bin. Both these strategies are tested for

classification of lesions into circumscribed/speculation/normal classes using MLP classifier. The results of the experiments using these strategies are noted in the Table 5.4.

Table 5.4 Effect of binning strategies on classification performance

Binning	Circumscribed			Speculation			Normal		
	TN rate (%)	TP rate (%)	Fscore (%)	TN rate (%)	TP rate (%)	Fscore (%)	TN rate (%)	TP rate (%)	Fscore (%)
Equal Size	95.0	67.0	72.7	90.0	73.0	68.3	86.0	91.0	90.4
Equal Freq	86.1	77.0	70.2	69.2	33.0	38.0	53.5	51.0	53.7

It was observed that for circumscribed, speculation as well as normal class the Fscore, Sensitivity (TP rate) and Specificity (TN rate) using equal size binning was better than that using equal frequency binning. Therefore, equal size binning strategy was adopted for rest of the experiments.

5.4 Comparison of Different Classifiers

Various classifiers can be used for classification of images. Three most commonly used classifiers (MLP, SVM, KNN) were used for classification of mammographic lesions into circumscribed, speculation and normal categories in order to compare the performances of different classifiers. The details of the method used can be found in Chapter 3. Optimum parameters used for the MLP, SVM and KNN classifiers are detailed in the Table 5.5.

Table 5.5 Optimum parameters for classifiers

MLP		**SVM**		**KNN**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
learning Rate (η)	0.3	Epsilon	$1.0\ e^{-12}$	K	6
Momentum (α)	0.2	Complexity constant (C)	1.0		
No.of iterations	500	No. of iterations	1000		
No. of Hidden Units	10	Tolerance	0.001		
folds for cross-validation	10	folds for cross-validation	10	folds for cross-validation	10

The results of circumscribed/speculation/normal classification using different classifiers for miniMIAS dataset are shown in the Table 5.6, Table 5.7 and results for LakeImaging dataset are shown in the Table 5.8, Table 5.9.

Table 5.6 Classification accuracy for finegrained classification (miniMIAS)

	Classification Accuracy (%) - miniMIAS			
Classifier	Circumscribed	Speculation	Normal	Overall
KNN	60.0	56.0	70.0	70.0
SVM	42.0	61.0	82.0	71.0
MLP	72.7	68.3	90.4	82.0

Table 5.7 Sensitivity and specificity for fine-grained classification (miniMIAS)

	Sensitivity (%)				Specificity (%)			
Classifier	Circ	Spec	Normal	Overall	Circ	Spec	Normal	Overall
KNN	66.7	57.9	75.4	70.0	88.2	88.9	72.1	79.1
SVM	33.3	63.2	89.5	69.2	92.1	90.1	65.1	76.3
MLP	66.7	73.7	91.2	82.0	94.7	90.1	86.0	88.9

Table 5.8 Classification accuracy for finegrained classification (LakeImaging)

	Classification Accuracy (%) - LakeImaging			
Classifier	Circumscribed	Speculation	Normal	Overall
KNN	66.7	70.0	91.6	79.8
SVM	66.7	71.4	96.0	82.0
MLP	80.0	82.0	94.0	84.0

Table 5.9 Sensitivity and specificity for finegrained classification (LakeImaging)

	Sensitivity (%)				Specificity (%)			
Classifier	Circ	Spec	Normal	Overall	Circ	Spec	Normal	Overall
KNN	65.0	66.7	95.0	80.2	90.2	91.7	87.8	89.4
SVM	65.0	71.4	97.5	82.7	90.2	90.0	90.2	91.4
MLP	70.0	71.4	97.5	84.0	93.4	91.7	95.1	94.6

Figure 5.7 shows graphical representation of three classifiers for miniMIAS as well as LakeImaging dataset. For unbalanced datasets, classification accuracy is not the perfect measure for classifier performance. Hence, sensitivity and specificity parameters along with the Fscore were considered for comparing the performances of the classifiers. Figure 5.7 shows the graphical representation of various parameters (Fscore, Sensitivity and Specificity) using three different classifiers (MLP, SVM and KNN) for miniMIAS as well as LakeImaging datasets. It is observed that the MLP classifier has performed better than the SVM and KNN classifiers with respect to Fscore, Sensitivity and Specificity for both the datasets.

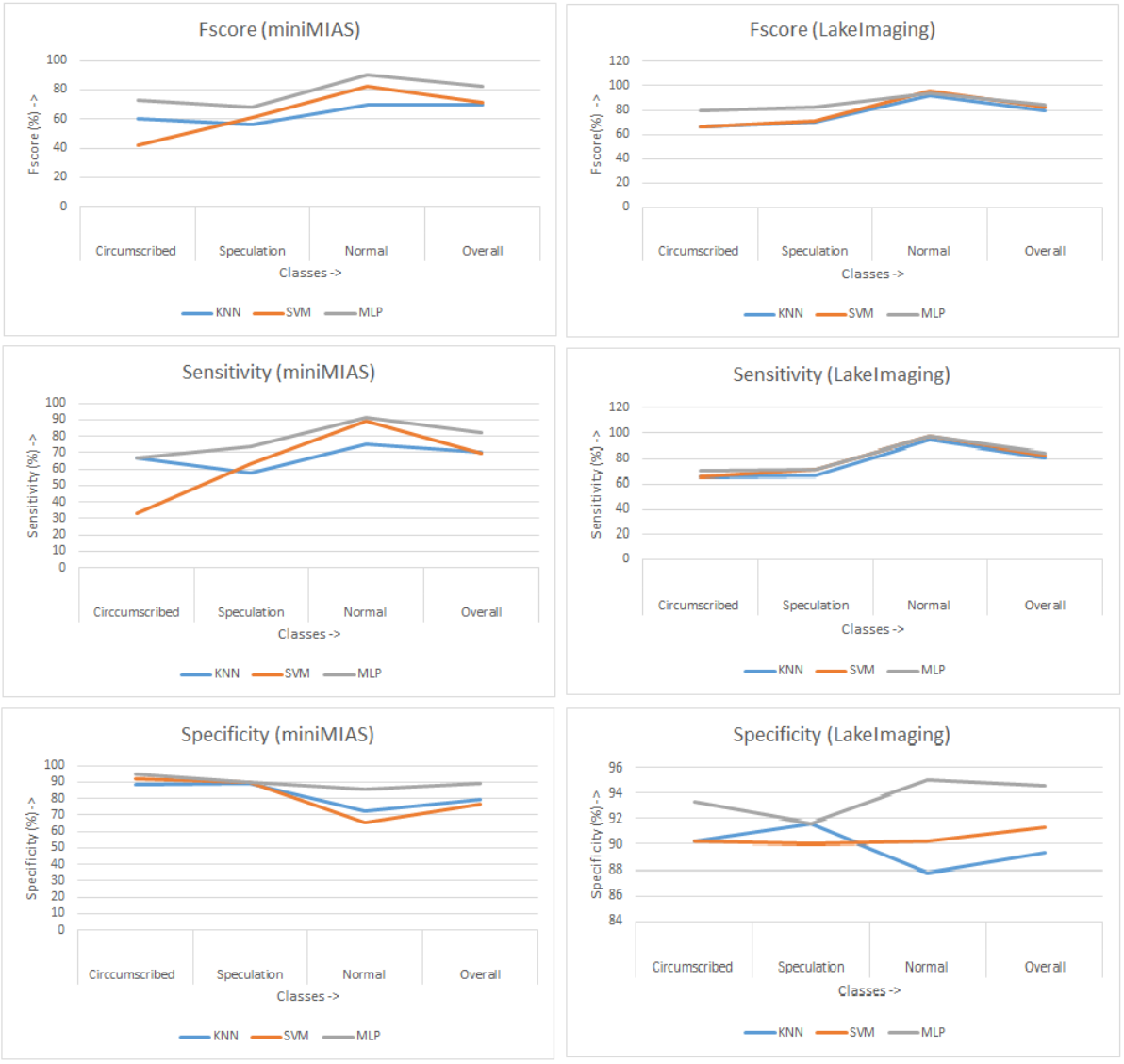


Figure 5.7 Classifier comparison using Fscore, sensitivity and specificity

Here parametric test (two-tailed paired T-test) is chosen in order to compare the performances of classifiers because the data is normally distributed and parametric tests are more sensitive and efficient as compared to the non-parametric tests. The results of the T-test for miniMIAS as well as LakeImaging datasets are given in the Table 5.10.

Table 5.10 T-test for classifier performance comparison

Dataset	**Classifiers used**	**p value**
miniMIAS	KNN and MLP	4.84627E-05
	SVM and MLP	0.001765148
LakeImaging	KNN and MLP	0.000616922
	SVM and MLP	0.016800785

The p-value for performance comparison between KNN and MLP, SVM and MLP for both the datasets (miniMIAS and LakeImaging) was found to be less than the level of significance α=0.05. Thus it is inferred that the MLP classifier performs significantly better than the KNN or SVM for mammographic classification using 3-gram features.

A true measure for comparison of classifier performances is Receiver Operating Characteristic (ROC) curve. This curve is plotted using the sensitivity/true positive rate on X axis and (1-specificity/true negative rate) on Y axis. Thus the trade-off between false negative and false positive errors can be achieved by analysing the ROC curve.

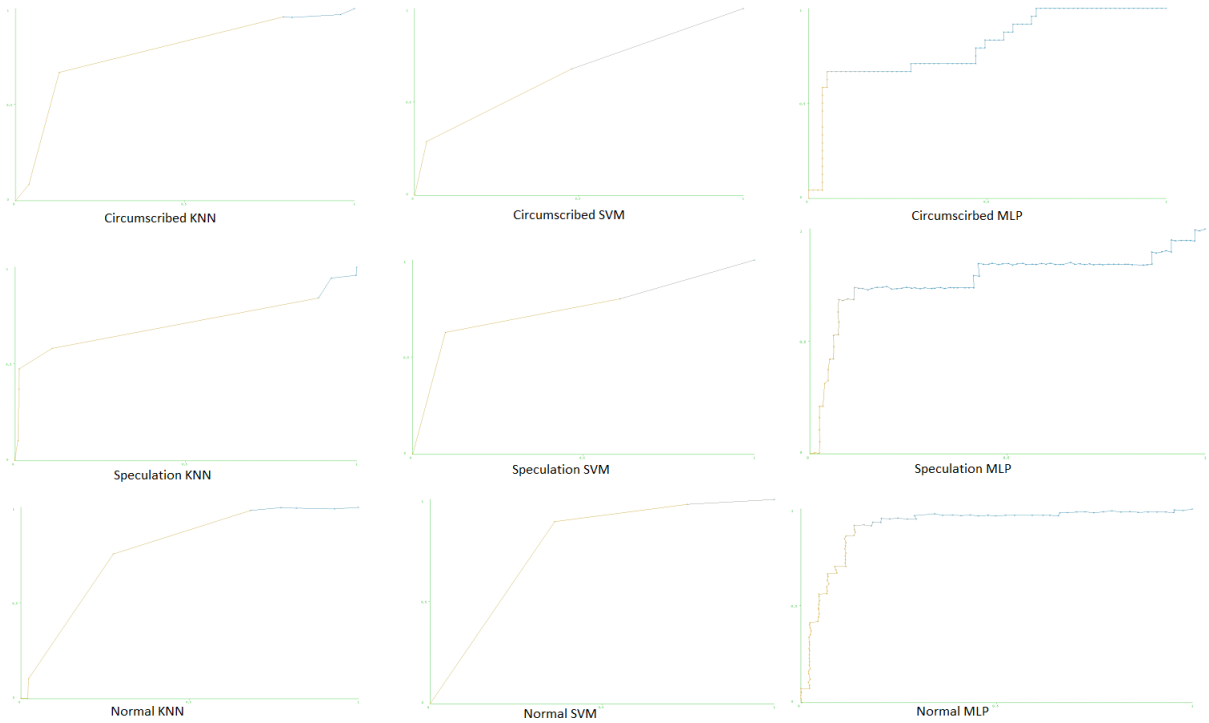


Figure 5.8 Receiver Operating Characteristics (ROC) curves for different classifiers

ROC curves for circumscribed, speculation and normal classes for miniMIAS dataset can be seen in the Figure 5.8. It is very clear from the ROC curves that for every class the MLP classifier performance is superior to the SVM or KNN classifier. Therefore, for all further experiments on miniMIAS and LakeImaging dataset MLP classifier is used.

5.5 Effect of Choosing Different Normalisation Techniques

Different types of normalisation strategies are used to fit the input data in the required range for getting better performance from the classifiers. Here three most widely used normalisation strategies (Min-Max, Zscore and Tf-idf) were tried to find out the effect on classification performance. As stated earlier, MLP classifier provided best performance and hence it was

used for this experiment. The experiments were carried out on miniMIAS dataset of mammography. Table 5.11 shows the result of using different normalisation techniques.

Table 5.11 Effect of different normalisation techniques

	Fscore (%)			
Normalization	Circumscribed	Speculation	Normal	Overall
Zscore	70.9	55.2	86.6	78.2
Min-Max scaling	72.7	68.3	90.4	82.0
TF/IDF	75.5	70.2	92.6	85.4

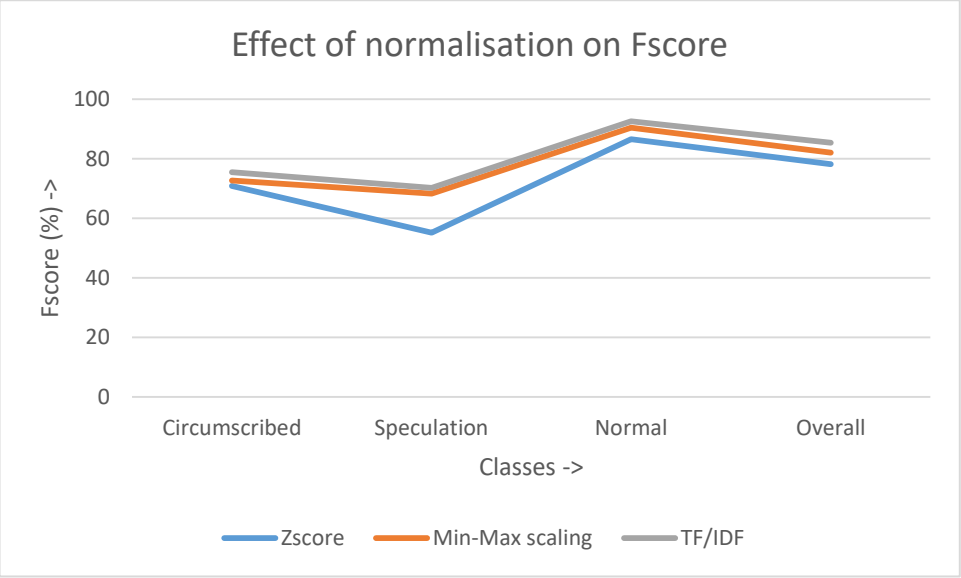


Figure 5.9 Effect of normalisation on Fscore

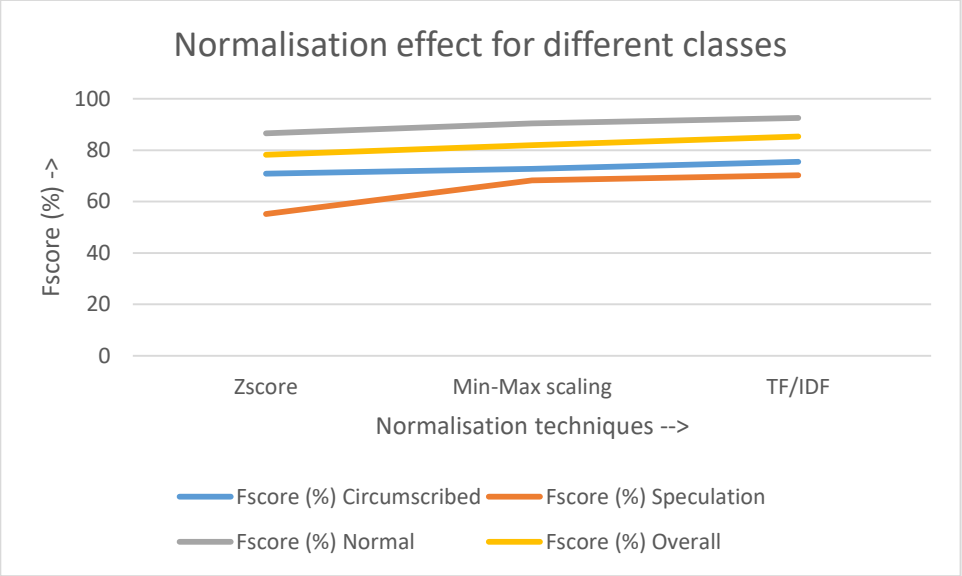


Figure 5.10 Fscore trend with various normalisation techniques

From Figure 5.10, it can be seen that Zscore offers worst performance especially for the speculation class. The Fscore for normal, circumscribed classes were slightly increased as the normalisation technique changed from zscore to min-max. On the other hand the Fscore for the speculation class was increased substantially with change in normalisation from zscore to min-max. Further, change in normalisation from min-max to tf-idf resulted in a slight increase in the Fscore for all the three classes. Thus tf-idf gives better performance than zscore and min-max normalisation. Therefore, this normalisation technique was chosen.

The experiments explained so far were aimed at determining the optimum configuration required for use of Pixel N-grams for mammographic classification. The optimum value of grey scale reduction was found to be 8 and the optimum value of N was found to be 3. Further, MLP classifier with tf-idf normalisation technique was found to give best performance. These optimum settings were used for all of the further experiments.

Next, the results of experiments on classification of lesions into normal/abnormal, and circumscribed/speculation/normal categories are noted and discussed. Further, feature selection and piecewise constant approximation experiments which try to improve the circumscribed/speculation/normal classification results are detailed.

5.6 Normal/Abnormal Classification

Normal/Abnormal classification is quite useful for automated screening of mammograms for breast cancer. The efficacy of 3-gram features for normal/abnormal classification was tested on miniMIAS and LakeImaging datasets. Three gram features were chosen for classification

as it was found to be the optimum value of N for both the datasets.TF/IDF normalisation was found to provide better results than the zscore or min-max normalisation. Hence, MLP classifier along with TF/IDF normalisation technique was used for classifying the ROIs into normal and abnormal categories.

In order to compare the performance with other existing techniques two techniques were chosen: Intensity histogram and Haralick features (Haralick et al., 1973) based on co-occurrence matrix. The software programs for computing Intensity histogram features and Haralick's features were implemented in Matlab. The intensity histogram features were computed for all the ROIs and used for normal/abnormal classification using MLP classifier. Similarly, Haralick's features were computed for every ROI and were used for normal/abnormal classification using MLP classifier. Results of normal/abnormal classification for miniMIAS dataset using Intensity histogram, Haralick's features and 3-gram features are noted in the Table 5.12. These results for miniMIAS dataset are shown graphically in Figure 5.11.

Table 5.12 Normal/Abnormal classification performance (miniMIAS)

Parameters	Intensity Histogram	Haralick features	3-gram features miniMIAS	3-gram features LakeImaging
Abnormal Accuracy (%)	41.5	76.19	80.95	95.1
Normal Accuracy (%)	80.8	88.23	91.18	95.0
Overall Accuracy (%)	60.8	83.63	87.27	95.06
Sensitivity (%)	39.5	76.19	80.95	95.01
Specificity (%)	78.5	88.23	91.17	95.1

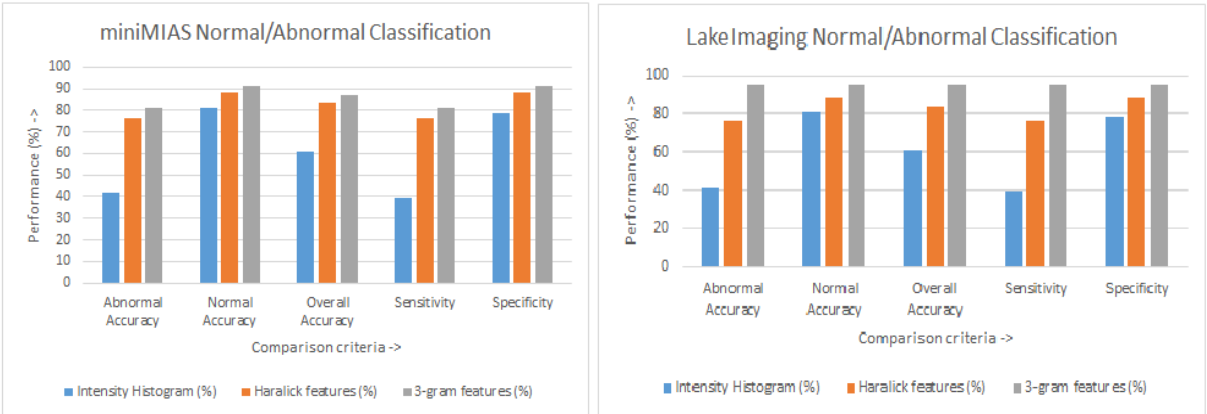


Figure 5.11 Comparison of different features (miniMIAS)

Comparison of classification performance using Pixel N-grams with existing techniques was carried out by using criteria such as classification Accuracy, Sensitivity and Specificity.

A two tailed paired T-test was performed using the classification performance parameters (abnormal accuracy, normal accuracy, overall accuracy, sensitivity and specificity) for both the datasets and noted in the Table 5.13 to see if the difference between classification using Intensity histogram and 3-gram features is statistically significant. Likewise, the T-test was also performed for comparing the classification performance using Haralick's features and 3-gram features for both datasets.

Table 5.13 T-test results for normal/abnormal classification (miniMIAS)

Dataset	**Feature sets used**	**p value**
miniMIAS	3-gram and Histogram	0.015
	3-gram and Co-occurrence matrix	0.000
LakeImaging	3-gram and Histogram	0.016
	3-gram and Co-occurrence matrix	0.009

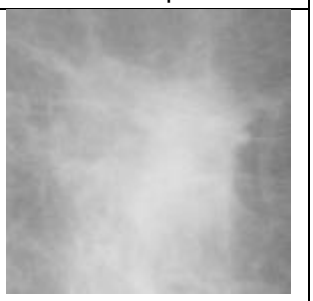
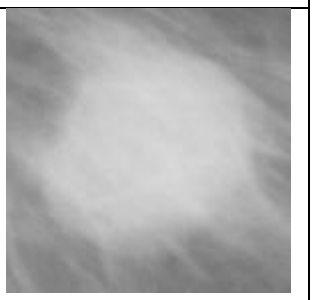
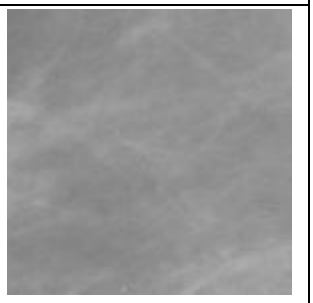
The p-value obtained after conducting T-test between classification performance using 3-gram and Histogram technique for normal/abnormal classification was found to be 0.015847 for miniMIAS and 0.016511 for LakeImaging dataset; below the level of significance $\alpha = 0.05$. Therefore, it can be concluded that the normal/abnormal classification using 3-gram features performs significantly better than the classification using histogram features. Similarly, the p-value calculated using the T-test between classification performance using 3-gram features and Haralick's features was found to be 0.000733 for miniMIAS and 0.00975 for LakeImaging dataset, which is also much less than the level of significance $\alpha = 0.05$. Therefore, it can be inferred that the classification performance using 3-gram features was significantly better than the classification performance using Haralick's features. Thus, for the miniMIAS dataset it can be concluded that the 3-gram features outperform the existing features such as histogram and Haralick's features with respect to accuracy, sensitivity as well as specificity for normal/abnormal classification.

Normal/abnormal classification is useful for the automated breast screening process; however, for diagnostic purposes fine-grained classification is necessary. In the next section, experimental results of classification based on various types of lesions (circumscribed, speculation/normal) for miniMIAS as well as LakeImaging dataset are noted.

5.7 Circumscribed/Speculation/Normal Classification

Shape and texture of the lesion plays an important role in diagnosing the breast lesion. Table 5.14 shows various types of lesions that appear on mammogram along with their possible diagnosis. A circumscribed lesion is the lesion whose margins are well defined meaning easily distinguishable. On the other hand speculated lesions contain fine white lines radiating out from the borders. Circumscribed lesions are mostly benign whereas, the speculated lesions are suspicious of malignancy (cancer). The classification of the lesions into three categories circumscribed/speculation/normal is thus very beneficial for automated diagnosis of breast cancer (Ferreira & Borges, 2003; Rangayyan et al., 2000).

Table 5.14 Different types of breast lesions and possible diagnosis

Finding	Possible Diagnosis	Example
Speculated lesion	Cancer Post operative scar Breast abcess Fat necrosis	
Round or oval circumscribed mass	Cysts Cancer Galactocele Dilated duct Papilloma Phylloides tumor Fibroadenoma	
Normal		

In this experiment the classification is aimed to fall into three categories: circumscribed, speculation and normal rather than the coarse grained abnormal-normal dichotomy used in many studies. The 3-gram features were used along with MLP classifier for classification of ROIs. The 3-gram features were normalised using tf-idf normalisation. The classification performance was measured using Sensitivity, Specificity, Fscore criteria. The results for the

classification can be seen in the Table 5.15 for miniMIAS dataset and Table 5.16 for the LakeImaging dataset respectively. In order to compare the performance of 3-gram features with other existing techniques Intensity histogram and Haralick's features were chosen. The histogram and Haralick's features were computed using Matlab programs. The generalisation was estimated using 10 fold cross-validation.

Table 5.15 Circumscribed/Speculation/Normal classification (miniMIAS)

Features ->		Histogram	Haralick	3-gram
Circumscribed	Fscore (%)	23.5	30.2	72.7
	Sensitivity (%)	20.0	30.5	67.0
	Specificity (%)	75.0	65.0	95.0
Speculation	Sensitivity (%)	50.0	60.3	73.0
	Specificity (%)	60.0	80.0	90.0
	Fscore (%)	43.5	60.5	68.3
Normal	Sensitivity (%)	50.0	60.0	91.0
	Specificity (%)	75.0	80.0	86.0
	Fscore (%)	50.0	60.3	90.4
Overall	Fscore (%)	40.0	50.0	82.0

Table 5.16 Circumscribed/Speculation/Normal classification (LakeImaging)

Features ->		Histogram	Haralick	3-gram
Circumscribed	Fscore (%)	55.0	63.8	80.0
	Sensitivity (%)	55.0	68.0	70.0
	Specificity (%)	85.2	85.2	90.16
Speculation	Fscore (%)	61.9	47.1	82.0
	Sensitivity (%)	61.9	38.1	71.4
	Specificity (%)	86.6	90.6	91.66
Normal	Fscore (%)	92.5	93.8	94.0
	Sensitivity (%)	92.5	95.0	97.5
	Specificity (%)	85.5	86.1	87.8
Overall	Fscore (%)	75.3	76	83.95

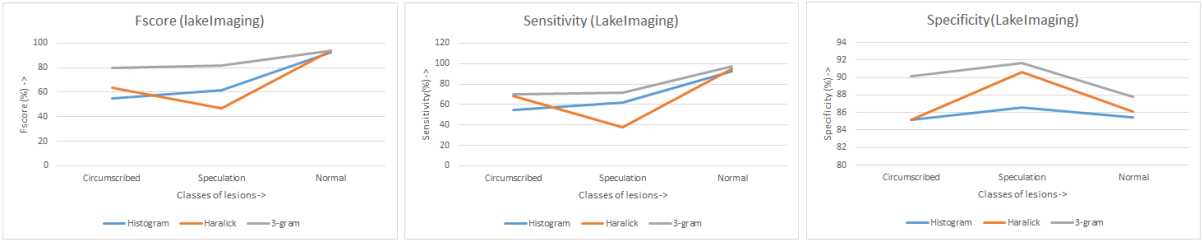


Figure 5.13 Circumscribed/speculation/normal classification (LakeImaging)

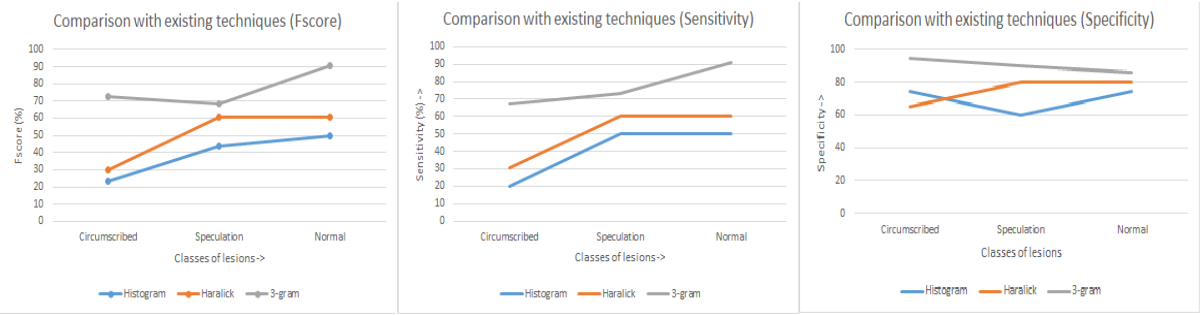


Figure 5.12 Circumscribed/speculation/normal classification (miniMIAS)

For miniMIAS dataset, Figure 5.12 provides the graphical representation of the classification results (Fscore, sensitivity and specificity) for Intensity histogram, Haralick's features and Pixel N-gram features. For circumscribed, speculated and normal classes the Fscore using histogram and Haralick's features is comparable but low whereas, the Fscore using Pixel N-gram features is considerably higher. Sensitivity with circumscribed class is very low using histogram and Haralick's features. The Fscore as well as sensitivity using Pixel N-gram features was high for all the three classes with normal class achieving the highest of all. Further, the specificity was also found to be quite high for all the classes (slightly less for normal class).

Similarly, the graphical representation of finegrained classification performance (Fscore, sensitivity and specificity) for LakeImaging dataset is shown in Figure 5.13. It can be seen that the Fscore for circumscribed and speculation class using Pixel N-gram features was much better than that using Intensity histogram and Haralick features whereas, the Fscore for normal class was found to be comparable with other two techniques. Sensitivity using Pixel N-gram features was observed to be much better than that using Intensity histogram and Haralick features for speculation class however, was comparable for circumscribed and normal class. Further, the specificity using Pixel N-gram features was high as compared to other two techniques for all the classes.

Thus the classification of mammographic lesions into circumscribed/speculated/normal categories using Pixel N-gram features achieved high values of Fscore, Sensitivity and Specificity for miniMIAS as well as LakeImaging dataset. The high values of Fscore, Sensitivity and Specificity are beneficial for clinical purposes and hence Pixel N-grams technique is a promising candidate for automated classification of breast lesions for diagnostic purposes.

A series of two tailed paired T-tests were conducted for comparing the classification performance using 3-gram features with the classification performance using Intensity histogram and classification performance using Haralick's features. Fscore, Sensitivity and Specificity values for circumscribed, speculation and normal classes were considered for the T-tests. While choosing the significance level α for a T-test it should be kept in mind that it is less likely to make Type I error and more likely to make a Type II error by choosing smaller value of α. Type I error is creating more false positives while a Type II error is failing to detect a lesion that is present (false negatives). However, in medical detection systems Type II errors are more serious. Table 5.17 shows T-test results for comparing the 3-gram features with existing techniques.

Table 5.17 T-test results for circumscribed/speculation/normal classification

Dataset	**Feature sets used**	**p value**
miniMIAS	3-gram and Histogram	0.0000
	3-gram and Haralick's features	0.0010
LakeImaging	3-gram and Histogram	0.0077
	3-gram and Haralick's features	0.0511

The p-values obtained after having t-tests for 3-gram and Intensity histogram classification results (miniMIAS and LakeImaging) were found to be quite less than the level of significance ($\alpha = 0.5$). Therefore, we could be sure that the classification using Pixel N-gram features provides significantly better performance than the classification using intensity histogram features. Similarly, the p-value obtained after running the t-test for the 3-gram and Haralick's features for miniMIAS dataset is below the level of significance ($\alpha = 0.5$). Thus the classification results using 3-gram features can be seen to be significantly better than that using Haralick's features for circumscribed/speculation/normal classification. On the other hand for LakeImaging dataset the p value obtained is slightly greater than the level of significance

indicating no significant performance difference. Thus it is demonstrated that the Pixel N-grams are better at distinguishing among various types of mammographic lesions and could be quite useful for automated diagnosis of breast cancer.

5.8 Feature Selection using Wrapper Approach

In case of classification, most of the times it is possible that some features are redundant or irrelevant. The redundant features cause more training time whereas the irrelevant features can result in overfitting thus less generalised classification model. The process for identifying the best subset of features from all the features are selected is known as feature selection. Feature selection is performed for three main reasons.

- Reduce overfitting and hence better generalisation
- Improve accuracy
- Reduce training time

Weka provides feature selection tool using the attribute evaluator and search method. The most commonly used feature selection methods are a) wrapper method b) filter method. Wrapper method (wrapper subset evaluator, weka) has been used here to find the most significant features. A wrapper subset evaluator creates all possible subsets of features from the feature vector. Then it uses the classifier algorithm using each subset. The subset of features with highest classification performance is declared as the best feature subset. Features selected using wrapper approach for circumscribed/speculation/normal classification for both the datasets (miniMIAS and LakeImaging) are noted in the Table 5.18. The average counts of these selected features for LakeImaging dataset are graphically shown in the Figure 5.14.

Table 5.18 Best 3-gram features selected using wrapper approach

Datasets ->	**miniMIAS**	**LakeImaging**
3-gram Features	1 1 1	1 1 2
	2 2 2	2 1 2
	3 3 4	3 5 5
	3 4 3	4 3 2
	4 5 4	5 3 3
	5 6 6	5 5 5
	6 7 7	8 2 2
	7 7 7	8 8 8

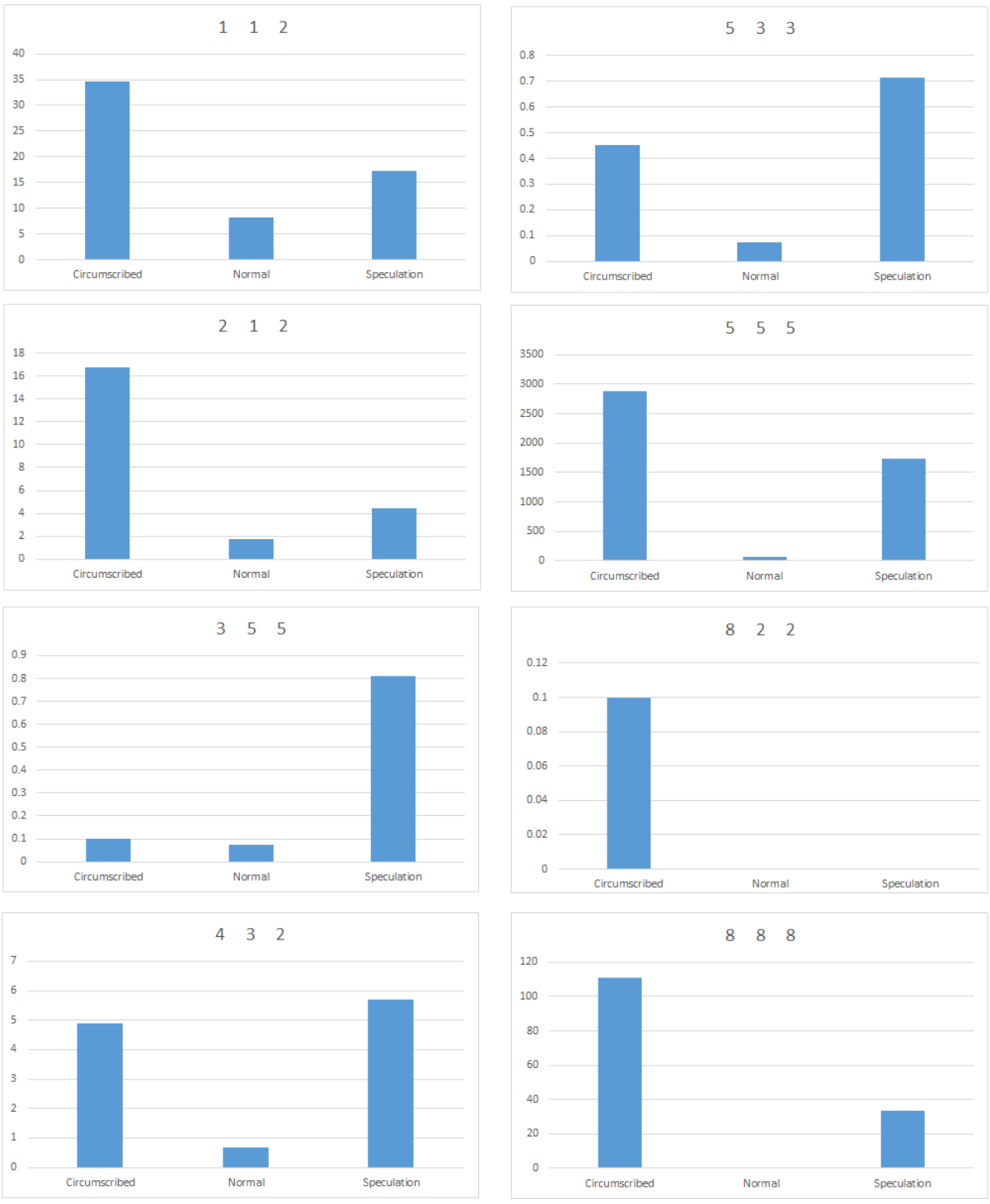


Figure 5.14 Average counts of best features for LakeImaging dataset

The classification performance for all the classes using the selected feature subset is noted in Table 5.19. It can be seen that the Fscore is improved with the feature selection for both the datasets. An overall Fscore of 85.5% for miniMIAS and 87.65% for LakeImaging dataset was obtained. Due to less number of features, the classification time is also reduced.

Table 5.19 Classification performance using best feature subset

Datasets	Fscore (%)			
	Circumscribed	Speculation	Normal	Overall
miniMIAS	81.3	80.4	91.0	85.5
LakeImaging	83.7	84.4	94.0	87.65

5.9 Piecewise Constant Approximation

Piecewise constant functions are functions for which the space can be subdivided into subspaces over each of which the function takes the same value. This would result in much less number of parameters or sets of coefficients which could then be used as features for classification of images. The need for piecewise constant approximation is explained in Chapter 3.

For this method N-gram features of every image were computed considering all the 256 grey levels. Thus a list of points (N-grams) and their associated function values (number of occurrences/counts) were obtained. Then every image was represented using the piecewise constant approximation.

However, it is a difficult problem in 3 dimensions. Therefore, an alternative way to achieve the piecewise constant approximation was used with the help of k-means clustering. Here, the 3-gram counts using 256 grey levels were considered. Then k-means clustering was applied on each image in the dataset.

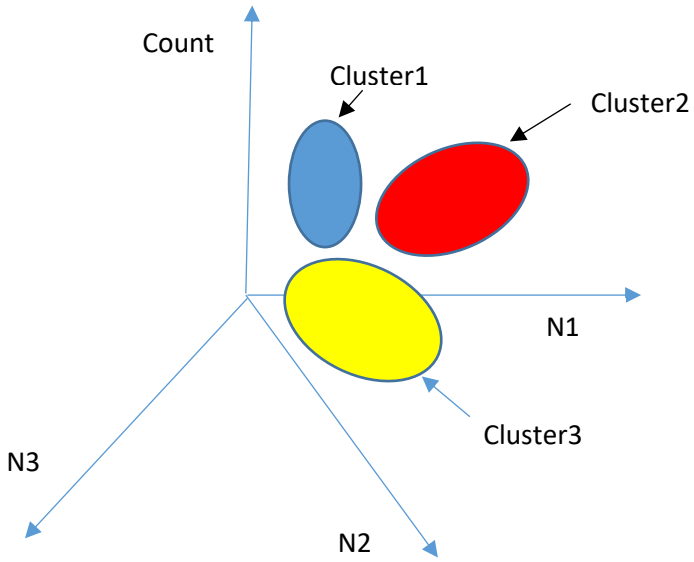


Figure 5.15 K-means clustering for piecewise constant approximation

The clustering process thus produced a set of clusters and associated 4-d cluster centres. The list of cluster centres (N1, N2, N3, count) along with the number of points in the cluster were then used as the features for the image classification. Figure 5.15 shows how k-means clustering can be applied in 4 dimensions (N1, N2, N3,Count) for piecewise constant approximation of 3-gram representation of an image. The cluster centres along with the number of points in the cluster were then used as input to the MLP classifier for classifying lesions into circumscribed/speculation/normal categories.

Results of the lesion classification using piecewise constant approximation features on miniMIAS dataset are given in the Table 5.20.

Table 5.20 Classification using piecewise constant approximation

	Circumscribed	Speculation	Normal
Fscore (%)	75.2	71.5	92.0
Sensitivity (%)	70.6	70.3	93.0
Specificity (%)	95.8	93.3	88.7

In the piecewise constant approximation all the 256 grey levels of pixels were considered and hence the problem of finding the optimum value of grey level is solved. Also, the clustering results in much less numbers of features and they are dependent upon the value of k used for the K-means algorithm. Various values of K were tried and the best performance was obtained with k = 6 and is noted in the Table 5.20. This representation provides slightly better classification performance as it produces a higher level representation of images. However, the classification results using piecewise constant approximation were not significantly better than the classification performance using 3-grams after grey scale reduction of images to 8 grey levels. This suggests that the grey level reduction value of 8 grey levels is the optimum value.

5.10 Computational Complexity Comparison

With mammographic classification, computational cost is one of the important factors needed for increasing the efficiency of the radiologists (Tourassi, Harrawood, Singh, & Lo, 2007; B. Zheng, 2009). One of the most widely used features for mammographic classification are Haralick's features (Haralick et al., 1973) based on co-occurrence matrix. The aim of this investigation is to compare the computational complexity of Pixel N-gram features with that of Haralick's feature computation.

Co-occurrence matrix is defined for an image to analyse the distribution of co-occurring pixel values at given offset and an angle Θ. Figure 5.16 shows the computation of co-occurrence matrix for a 4 × 4 image with 4 grey scales (0, 1, 2, 3).

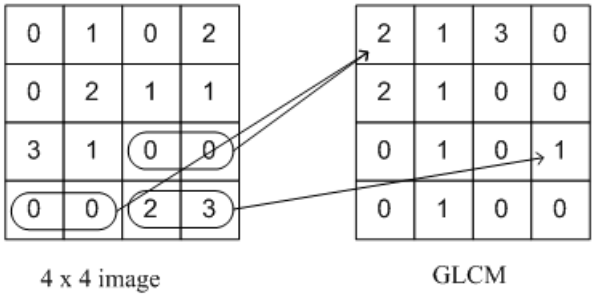


Figure 5.16 Co-occurrence matrix computation for 4 grey level image

There are 14 second order statistical features based on co-occurrence matrix defined by Haralick. The Haralick's features and the equations (Haralick et al., 1973) for computing them are given in the Table 2.1.

In order to exploit the computational complexity of the N-gram features experimentally, the time required to compute the 3-gram features, Intensity histogram features, Haralick's features was measured using standard tic and toc functions available in Matlab. For this experiment ROIs of 4 different sizes were extracted from every mammogram image of miniMIAS dataset. The time required to compute 3-gram features was noted for all the ROIs of size 70×70, 140×140, 280×280, 560×560. Then the average time required is calculated for the particular image size. Similarly, time required for the computation of Intensity histogram features and Haralick features was computed for all the ROIs. The average time required for computation of Intensity histogram, Haralick features and 3-gram features for different ROI sizes is noted in the Table 5.21.

Table 5.21 Computation time requirement for different features

Features	Computation time in milliseconds			
ROI size (Pixels)-->	70×70	140×140	280×280	560×560
Histogram	1.4	4.00	6.9	8.6
Haralick	12.44	73.98	119.62	235.42
3-gram	1.97	6.91	25.5	51.62

Figure 5.17 shows how the computational time requirement varies with respect to image size in pixels for Intensity histogram, Haralick features and 3-gram features. It was observed from this graph that the computation of 3-gram features is about 7 times faster than the computation of Haralick's features. Moreover, the increase in the computation time for 3-gram features is almost linear with respect to image size whereas, the increase in the computation time for

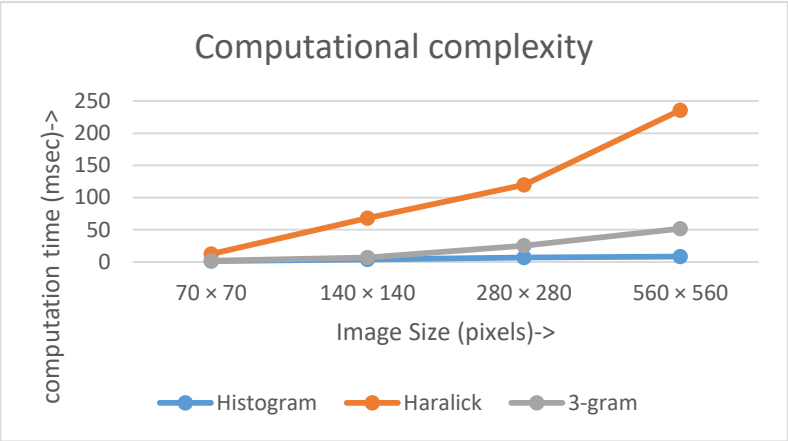


Figure 5.17 Computational time requirement for various features

Haralick's features appears to be exponential. On the other hand 3-gram computation time is comparable with that of the Intensity histogram features. The computational complexity of N-grams is thus demonstrated to be much less than Haralick's features.

5.11 Chapter Summary

In this chapter the use of Pixel N-gram features for mammographic classification was exploited. Various experiments were carried out on two datasets. The benchmark miniMIAS dataset with secondary digital images (x-rays mammograms digitized with the help of scanner) was used so that the performance of Pixel N-grams could be compared with the existing techniques. Secondly, a dataset specially prepared for this research project, provided by LakeImaging (primary digital images) was used to test the performance of Pixel N-grams on high resolution truly digital images.

Grey scale reduction of images was performed for both datasets so as to reduce the computational complexity and dimension of the feature vector. However, the decision of choosing the optimum value of grey scale reduction was a challenging task. It was observed that after grey scale reduction, all the possible numbers of N-grams were not present in the corpus. The proportion of the possible number of N-grams that were observed was found to decrease exponentially (Refer Figure 5.3) with the increase in number of grey scales used. It can be argued that a classifier can generalise well if the features used represent a higher proportion of the possible feature set. Therefore, a grey level where the proportion of possible N-grams which are actually observed begins to drop dramatically was chosen as the best grey scale reduction value which is observed to be 8.

Further, grey scale reduction is nothing but discretisation of the intensity values of pixels in the images. Two types of discretisation strategies are widely used: equal size binning and equal frequency binning. Classification performance (circumscribed/speculation/normal) with both the types of binning strategies using 8 bins was analysed. The performance with equal size binning strategy was found to be superior to the performance with equal frequency binning strategy. Thus equal size binning strategy with 8 grey level bins was used for all the classification experiments.

Then it was necessary to analyse the effect of varying N on the classification performance and finding out optimum value of N where the performance is best. Values of N = 1, 2, 3, 4 and 5 were used for this experiment. Although, it is arguable that with increase in N, image representation becomes more and more complete thus resulting in better classification performance, the actual observation was quite different. For both the datasets (miniMIAS and LakeImaging), it was observed that the classification performance (circumscribed/speculation/normal) increased as N was increased from 1 to 3, however started dropping with further increase in N (4, 5). Thus the best classification performance was obtained at N = 3 which can be declared as the optimum value of N. One of the possible reasons for this could be as N is increased the longer sequences become hard to find resulting in sparse feature vector. Also, these longer sequences then become too specific for a particular image making it hard for a classifier to generalise. Further, for mammographic lesion classification the features are required to distinguish between the different shapes, boundaries, and textures. It seems that 3-gram features are quite efficient in distinguishing between various shapes and boundaries. Thus, the features with N greater than three do not seem to add a lot of useful information for classification purposes.

Different classifiers work differently on the input data. Three most widely used classifiers (MLP, SVM, KNN) were tested for classification of lesions into circumscribed, speculation and normal categories. Classification performance of these three classifiers was compared using Fscore, Sensitivity and Specificity criteria. Two tailed paired T-tests were performed to compare the performances and it was observed that the p-values for both the datasets (miniMIAS, LakeImaging) were less than the level of significance $\alpha = 0.05$. Further, Receiver Operating Characteristic (ROC) curves were plotted for each class using different classifiers. It was observed that the ROC curves with MLP classifier were the best for circumscribed, speculation as well as normal classes. Thus, it can be concluded that the classification

performance using MLP classifier was significantly better than that using SVM or KNN classifiers. Hence, a MLP classifier was chosen for the rest of the experiments.

The classifier usually performs better if the input values are in a particular range. Various normalisation techniques are in place to convert the input values into the required range. Z-score, Min-Max and Tf-idf techniques were analysed in this study. The classification performance with Tf-idf normalization was found to be better than the Z-score and Min-Max normalisation.

The optimum settings found empirically (grey scale reduction to 8 grey levels using equal size binning, N = 3, MLP classifier, Tf-idf normalization) were then used for classification of mammographic lesions. Two types of classifications were observed. Normal/abnormal classification which is useful for automated screening purpose and circumscribed/speculation/normal classification useful for automated diagnosis of the lesions. The classification performance was tested on miniMIAS as well as LakeImaging dataset. In case of normal/abnormal classification for miniMIAS dataset; the Fscore was noted as 80.9% for abnormal class whereas, 91.1% for normal class with overall Fscore of **87.27%.** The experiments on LakeImaging denote the Fscore of 95.1% for abnormal class and 95% for normal class with an overall Fscore of **95.06%.**

The miniMIAS is a benchmark database for mammographic research. Hence a lot of work has been done by many researchers on this dataset. Table 5.22 summarizes the results of some of the works for normal/abnormal classification on miniMIAS dataset.

Table 5.22 Normal/abnormal classification results by various researchers (miniMIAS)

Work	Features	Classifier	Validation	Accuracy	Sensitivity	Specificity
Christoyianni 1999	GLCM	MLP	No	82.35	-	-
Bovis 2000	Haralick+ Inertia + Diff Avg.	ANN	10 fold	77	74	-
Mousa 2005	Wavelet	Adaptive neurofuzzy inference	No	81.4	-	-
Khuzi 2009	GLCM	K-means	No	82	72	88
Proposed	Pixel N-grams	MLP	10 fold	87.27	80.95	91.17

The number of parameters considered for classification in various studies such as number of images used, features used, classifiers used, ROI size used, validation used to estimate the generalisation differ significantly and hence direct comparison of these results with proposed technique (Pixel N-grams) is not quite possible. However, the results using Pixel N-grams seem to be better than these existing techniques.

In order to directly compare the results, the Intensity histogram as well as Haralick's feature computation algorithms were implemented using Matlab and the classification of lesions using these features were also tested using Fscore, Sensitivity and Specificity. Overall Fscore of 60.8% was achieved using Intensity histogram technique whereas 83.63% was achieved using Haralick's features for miniMIAS dataset. A T-test was performed for comparing the performances and it has been observed that the normal/abnormal classification using Pixel N-gram features is significantly better than that of Intensity histogram as well as Haralick's features.

In case of fine grained classification (circumscribed/speculation/normal) the Fscore obtained for miniMIAS dataset was 72.7% for circumscribed class, 68.3% for speculation class and 90.1% for normal class with an overall Fscore of **82%**. For the LakeImaging dataset on the other hand, the Fscore obtained was 80% for circumscribed, 82% for speculation and 94% for normal class with overall Fscore of **87.65%.** Thus it is evident that Pixel N-grams can be used for normal/abnormal classification and therefore, quite useful for automated screening of breast cancer. Further, the Fscores obtained for all the classes for LakeImaging dataset are higher than 80% however, the Fscores for miniMIAS dataset are less than 80%.

Some of the misclassified instances from miniMIAS dataset are given in the Figure 5.18. It can be seen from the Figure 5.18 that the normal ROIs which were misclassified as circumscribed or speculation does seem to have some kind of abnormality appearance.

Thus labelling these instances as abnormal and pointing these out for radiologists attention (false positives) was actually a good decision. Further, the instances which are circumscribed but classified as speculated do have some spicules radiating. On the other hand the lesions which are actually circumscribed but classified as speculated does not have very well defined boundaries. However, the false negatives are the ones which are more dangerous.

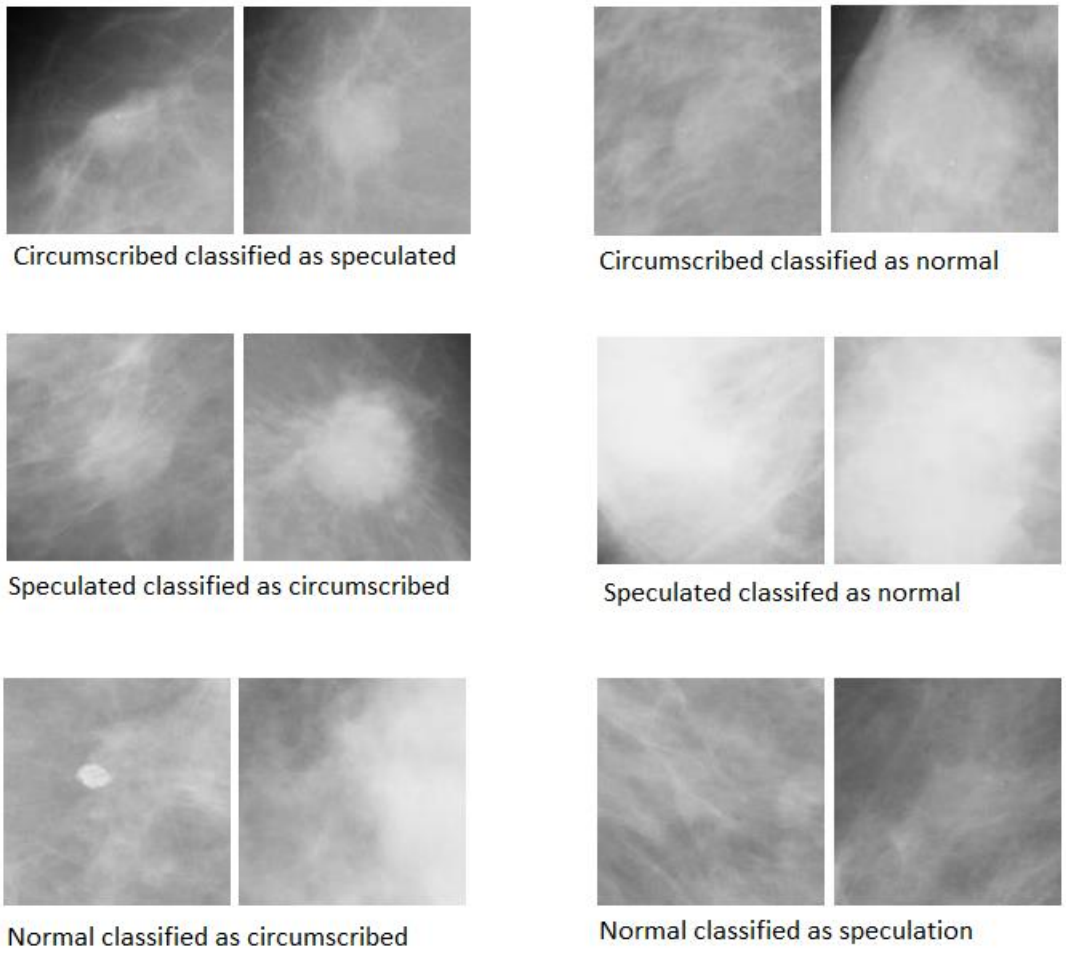


Figure 5.18 Misclassified instances from miniMIAS dataset

The background tissue of the breast plays an important role here. There are three types of background tissue which can be seen in women of varying ages. The three background tissue types are fatty, fatty-glandular and dense-glandular. The Figure 5.19 shows some examples of normal and abnormal ROIS with different types of background tissue. From the Figure 5.19 it is very obvious that the abnormality detection from the dense-glandular background tissue is quite difficult task.

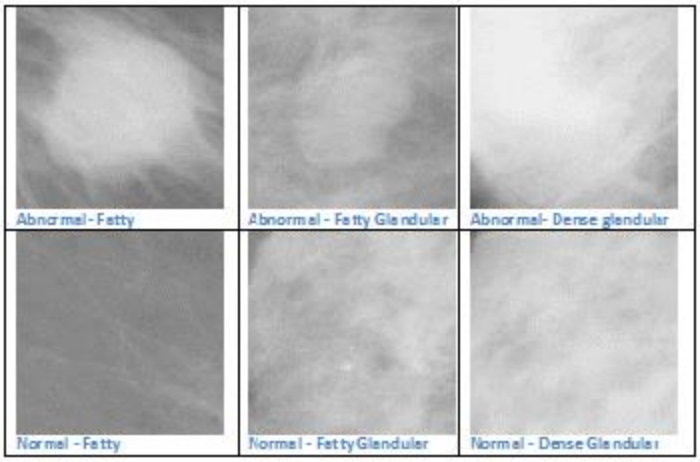


Figure 5.19 Different background tissue density

The lesions which are either circumscribed or speculated but classified as normal are the ones to be concerned about. It was observed that this happened mainly in case of two situations. These lesions were either embedded in the dense-glandular background tissue or were not fully enclosed in the ROIs.

It is clear from the observations that the Fscores obtained for LakeImaging dataset were greater than those for the miniMIAS dataset. This could easily be explained with the fact that the images from LakeImaging dataset are high resolution primary digital images and are of good quality whereas, the miniMIAS dataset images are secondary digital images and have low contrast. Thus, it is obvious that better resolution and quality of images results in better classification performance.

An effort to improve the classification performance was made by using the wrapper approach to select the best features. The Fscore obtained for miniMIAS dataset by applying feature selection was 81.3% for circumscribed, 80.45 for speculation and 91% for normal class with an overall Fscore of 85.5%. The Fscore for LakeImaging dataset after feature selection was noted as 83.7% for circumscribed, 84.4% for speculation and 94% for normal class with an overall Fscore of 87.65%. Thus it is demonstrated that with the best features selected Pixel N-grams are useful for circumscribed/speculation/normal classification and hence for automated diagnosis of breast cancer with an added advantage of low computation time.

To improve the performance further and to eliminate the need for deciding the optimum value of grey scale reduction, piecewise constant approximation was used. Three grams with all the

256 grey levels were considered for every image. These 3-gram features were clustered using k-means clustering producing the centroids in 4 dimensional space. The co-ordinates of the centroids (N1, N2, N3) along with the count and number of points in the cluster were then used as the input parameters to the classifier. The classification results obtained using piecewise constant approximation were slightly better than that of 3-grams using grey scale reduction to 8 grey levels. This observation was totally opposite to the expectation that this representation would greatly improve the performance. This could however mean that our selection of optimum grey level of 8 was the right decision.

The circumscribed/speculation/normal classification performance using Pixel N-grams was compared with that of Intensity histogram and Haralick's features using classification Accuracy, Sensitivity and Specificity. A series of two tailed paired T-tests were performed in order to compare the performance with Intensity histogram and Haralick's features. In medical detection or diagnosis systems type II errors are more dangerous. The p-values obtained for both the datasets (miniMIAS and LakeImaging) were found to be below the level of significance $\alpha = 0.05$. Therefore, it can be concluded that the classification performance for circumscribed/speculation/normal classification using Pixel N-grams was significantly better than that of Intensity histogram as well as Harlalick's features. Thus, Pixel N-gram features have been observed to outperform the existing Intensity histogram and Haralick features for classification of mammographic lesions.

In addition to the classification performance (Fscore, sensitivity, specificity), computational complexity of an algorithm to compute the features was also used for comparing Pixel N-gram features with existing feature computation techniques such as Intensity histogram and Haralick's features. This was done by measuring the time required to compute the various features using different sizes of ROIs. It was found that the Pixel N-gram feature computation was almost seven times faster than the Haralick's feature computation; whereas, the computational time required was comparable with the Intensity histogram features. The computational time for Haralick's features increased exponentially with increase in image size. On the other hand, the increase in computational time for Pixel N-gram features was linear with respect to increase in the image size. This is definitely a valuable feature for mammographic classification as this means low memory as well as processor requirements for diagnosis/detection purposes. This feature opens up a door to the accurate breast cancer

detection/diagnosis using low resource devices such as ipads, palmtops and mobile phones providing more convenience to the clinicians and patients.

6 Conclusion and Future Work

In this chapter, the results of the various experiments are summarised and related back to the research questions. The conclusions are drawn from the experimental results and are discussed thereafter. Then the limitations of the current research are discussed and the ideas for extending the Pixel N-grams technique and various possible applications are elaborated in the limitations and future work section.

6.1 Conclusion

In this thesis a novel feature extraction algorithm 'Pixel N-grams', inspired from the character N-gram concept in text retrieval is proposed for mammographic image classification. Two types of classification for mammograms were observed. Normal/abnormal classification is useful for automated breast cancer detection. On the other hand circumscribed/speculation/normal classification of lesions is useful for breast cancer diagnosis. The classification performance was measured using Fscore, classification accuracy, Sensitivity and Specificity. In case of cancer detection or diagnosis, false positives increase the patient's anxiety and health care costs whereas, false negatives can be fatal. Thus high values of sensitivity and specificity along with high values of classification accuracies are desirable.

The performance of mammographic classification using Pixel N-gram features was assessed using two different mammographic datasets. Benchmark miniMIAS dataset of mammography is publically available and contains secondary digital mammograms of equal resolution. On the other hand, the LakeImaging dataset contains primary digital mammograms of varying resolutions annotated specially for this project by experienced radiologists. Experiments on these mammographic databases include 1) Finding the optimum value of N, 2) Comparison of classifiers, 3) Effect of choosing different normalisation techniques, 4) Comparison of Pixel N-grams with existing feature extraction techniques (Classification performance and computation time), 5) Feature selection using wrapper approach and 6) Piecewise constant approximation.

Experiments revealed that mammograms classification performance increased with N up to 3 and then started dropping for miniMIAS as well as Lakeimaging dataset. Thus preferred value of N for classification of lesions for these databases was found to be 3. In an attempt to reduce the computational cost and effect of noise, grey scale reduction of images as a pre-processing

step was proposed. There is a trade-off between wanting to decrease grey levels so that the computational complexity is reduced and wanting to increase grey levels so that the features are not too coarse to diminish the classification accuracy. The optimum value of grey scale reduction was found to be 8 and hence images were grey scale reduced using 8 grey levels. Further, MLP classifier provided better Fscore, Sensitivity and Specificity than SVM or KNN classifiers for normal/abnormal as well as circumscribed/speculation/normal classification. Three different normalization strategies were utilized namely, Z-score normalization, Min-Max normalization and tf-idf weighting. Tf-idf normalization resulted in slightly higher classification accuracy than the Min-Max normalization, however min-max normalization was found to be the easiest and fastest normalization strategy.

The mammogram classification results were compared with existing techniques such as Intensity histogram and Haralick's features. It is evident that the Pixel N-gram features work significantly better than the Intensity histogram and Haralick features with respect to Fscore, Sensitivity (true positive rate) and Specificity for mammographic lesion classification. This is suspected to be due to the fact that Pixel N-gram features are able to represent local variations in an image more effectively. Further, Pixel N-grams are found to be good at normal/abnormal as well as circumscribed/speculation/normal classification for both the primary as well as secondary digital images.

Further, texture/density is an important characteristic of mammographic lesions. Therefore, texture image classification using Pixel N-grams was studied. It was evident that Pixel N-gram features can clearly distinguish between various texture images very well. It was found that the performance of Pixel N-grams for texture image classification was significantly better than that using existing statistical techniques such as Intensity histogram and Haralick's features. Also, texture classification performance using Pixel N-gram features was comparable with the texture classification using state-of-the-art BoVW technique (SIFT features).

Shape is another important characteristic of breast lesions. Hence basic shape classification using Pixel N-gram features was analysed in this work. Further, lesions can be of different sizes and located at different locations in a mammogram. Also, the mammographic images taken at different hospitals have varying resolutions. Experiments were carried out in order to discover the extent to which shape classification using Pixel N-grams is invariant to size, location of shape in the image and image resolution. It was observed that the Pixel N-grams

can classify various shapes irrespective of the size of the shape, location of the shape in an image as well as resolution of the image. Additionally, the shape classification results using Pixel N-grams were compared with the Intensity histogram and Haralick's features. Results demonstrate that the Pixel N-grams provide better performance than Intensity histogram and Haralick's features for shape classification.

Along with improving the accuracy, sensitivity and specificity, one of the aims of this research was to design a computationally efficient algorithm for classification of mammograms. Comparison of computational time for Pixel N-gram features with the Intensity histogram and Haralick's features was empirically carried out. It was demonstrated that the Pixel N-gram feature computation algorithms are simple to implement and were found to be almost seven times faster than the Haralick's feature computation. Further, the computational time for Pixel N-gram features varies linearly with respect to change in the image size.

From the current research, it can be concluded that the novel Pixel N-gram features are quite effective in distinguishing various textures and shapes. The Pixel N-grams perform significantly better than histogram and Haralick's features for texture, shape and mammograms classification. Further, they perform equally well as compared to the state-of-the-art bag of visual words features with an added advantage of simplicity and less computational cost. Another advantage of using Pixel N-grams features is that they do not require segmentation of the lesions and thus are not dependent upon the segmentation algorithm accuracy. Thus, the use of Pixel N-grams for mammographic lesion classification was found to be quite promising for breast cancer detection as well as diagnosis applications.

6.2 Limitations and Future Work

The proposed system works with an assumption that the lesion is completely enclosed in the extracted ROI. If this is not the case then the classification performance may diminish. Further, the methods requires ROI annotation from experienced radiologist and relies on the accuracy of the annotation. Also, the number of images used for the experiments were not very high due to the fact that every lesion has to be manually annotated by the radiologist. It would be interesting to analyse the performance of Pixel N-grams technique on a larger size dataset of mammography by applying automatic segmentation algorithm for ROI extraction phase. Further, the optimum number of N could be different for different datasets. Therefore, optimum

value needs to be determined for a new dataset in order to ensure good classification performance.

Texture is a variation of intensities quantified using measures such as smoothness, coarseness and regularity. The texture can be an important property useful for classification of various medical as well as non-medical images. Spectral features describe the global periodicity of grey levels by looking at high energy peaks. Similarly, statistical GLCM features model the spatial relationships amongst pixels by constructing matrices with different offsets and directions. Pixel N-grams proposed in this PhD project go one step ahead in modelling the texture by considering N consecutive grey level pixels. However, more complex patterns can be better described by considering the N neighbouring pixels separated by distance (offset) of d. Thus the Pixel N-gram algorithm could be enhanced to utilise the distance of 2, 3, 4 between the pixels to be considered in order to explore the classification of more complex texture patterns. Further, shape classification can be quite useful for object recognition applications in medical as well as non-medical image domain. In the work described in this thesis, performance of Pixel N-gram features for basic shapes (circle, square and triangle) using binary images was evaluated. However, the objects in the real world images are composed of complex shapes embedded in complex backgrounds. Also, these real world images are not binary images but they contain various grey levels. It is worthwhile studying the classification performance for more complex shapes where the images contain various grey level pixels. Thus, it would be interesting to see how Pixel N-grams work for complex shape classification where the shapes are embedded in complex backgrounds.

Mammographic lesions have texture and shape as the two important distinguishing characteristics. In this study it has been demonstrated that Pixel N-gram features are able to distinguish various textures and shapes and hence are found to be quite effective for mammographic lesion classification. Various other medical image classification tasks also depend upon features which are able to model textures and shapes. Therefore, the Pixel N-gram feature extraction technique could be easily adapted for classification of other medical images. Thus future work will involve exploring the extent to which Pixel N-grams are effective for other medical image classification tasks outside mammogram classification such as diabetic retinopathy, histopathological image classification, lung tumour detection using CT images, brain tumour detection using MRI images, wound image classification and tooth decay classification using dentistry x-ray images. Further, texture and shape classification is really

useful for classification of real world images outside the medical domain. Therefore, Pixel N-gram technique could be easily extended for applications such as object detection tasks for example, detection of car from the traffic videos, suspected thief detection from the surveillance video, motion detection and satellite image classification.

Further, these features are most likely to be useful for content based image retrieval (CBIR) tasks. In CBIR applications, a query image is given and the task is to retrieve images having similar features. Pixel N-grams describe the local image patterns as well as provide global intensity distribution. Hence, these features could be used to calculate the similarity between query image and images from the database. Such image retrieval applications will be helpful for medical research and training purposes. Furthermore, non-medical image retrieval applications could be benefited by this novel technique.

Medical images are vital part of diagnostics and patient treatment. With the advent of technology, there is a rapid increase in the number of radiological images produced every day. With existing storage mechanisms the images are stored in Picture Archiving and Communications System (PACS) servers whereas text such as patient history, radiology reports are stored in separate systems such as Hospital Information System (HIS) and Radiology Information System (RIS). While learning, training and clinical practice a radiologist stores the data about the past cases seen in textbooks and clinical practice in her/his memory. He or she then recalls this data in order to make a clinical decision about a new lesion (Muramatsu et al., 2005). Case Based Reasoning (CBR) suits well in clinical decision making as it comes with an appropriate explanation and justification of the solutions represented by the past cases and their similarities to the current case (Welter et al., 2011). With increasing amount of radiology case data, it is difficult to memorize and recall past similar cases. Moreover, for less experienced radiologists this practical data of past cases is not available in memory. Therefore, a facility to search for past similar cases is beneficial in diagnostic decision making. Further, there is some connection between the words/characters in the radiology report case and the features of lesions observed in the corresponding image for a case. One of the parts of this type of diagnostic system is image retrieval. As Pixel N-gram features have been proven to be successful for classification of mammographic images, they could also be used for similar image retrieval applications. The text based report retrieval could be achieved using N-gram text features. Thus, the similar case retrieval can be implemented by combining the Pixel N-gram image features and N-gram radiology report text features. It is surmised that this

combination of text and image N-grams will provide promising case retrieval results and will be explored in the future.

References

(RSNA), R. S. o. N. A. (2016, 09-2016). Mammography. http://www.radiologyinfo.org/en/info.cfm?pg=mammo.

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications, 46*, 139-144.

Aggarwal, N., & Agrawal, R. (2012). First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing, 03*(02).

Akgül, C. B., Rubin, D. L., Napel, S., Beaulieu, C. F., Greenspan, H., & Acar, B. (2011). Content-based image retrieval in radiology: current status and future directions. *Journal of digital imaging, 24*(2), 208-222.

Aksoy, S., & Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters, 22*(5), 563-582.

Alto, H., Rangayyan, R. M., & Desautels, J. L. (2005). Content-based retrieval and analysis of mammographic masses. *Journal of Electronic Imaging, 14*(2), 023016-023016-023017.

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2015). *Convolutional neural networks for mammography mass lesion classification.* Paper presented at the Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE.

Avni, U., Goldberger, J., Sharon, M., Konen, E., & Greenspan, H. (2010). *Chest x-ray characterization: from organ identification to pathology categorization.* Paper presented at the Proceedings of the international conference on Multimedia information retrieval.

Bankman, I. (2008). *Handbook of medical image processing and analysis*: academic press.

Bauckhage, C., & Tsotsos, J. K. (2005). *Bounding box splitting for robust shape classification.* Paper presented at the ICIP (2).

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and image understanding, 110*(3), 346-359.

Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-scale kernel machines, 34*(5).

Birdwell, R. L., Bandodkar, P., & Ikeda, D. M. (2005). Computer-aided detection with screening mammography in a university hospital setting 1. *Radiology, 236*(2), 451-457.

Bleyer, A., & Welch, H. G. (2012). Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine, 367*(21), 1998-2005.

Boggis, C. R., & Astley, S. M. (2000). Computer-assisted mammographic imaging. *Breast Cancer Research, 2*(6), 392.

Boiman, O., Shechtman, E., & Irani, M. (2008). *In defense of nearest-neighbor based image classification.* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.

Bouachir, W., Kardouchi, M., & Belacel, N. (2009). Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation. 215-220. doi:10.1109/sitis.2009.43

Bovis, K., & Singh, S. (2000). *Detection of masses in mammograms using texture features.* Paper presented at the Pattern Recognition, 2000. Proceedings. 15th International Conference on.

Brady, M., & Xie, Z.-Y. (1996). Feature selection for texture segmentation. *Advances in Image Understanding*, 29-44.

Brodatz, P., & Textures, A. (2009). A photographic album for artists and designers. 1966. *Images downloaded in July*.

Buist, D. S., Porter, P. L., Lehman, C., Taplin, S. H., & White, E. (2004). Factors contributing to mammography failure in women aged 40–49 years. *Journal of the National Cancer Institute, 96*(19), 1432-1440.

Caicedo, J. C., Cruz, A., & Gonzalez, F. A. (2009). Histopathology image classification using bag of features and kernel functions *Artificial intelligence in medicine* (pp. 126-135): Springer.

Carneiro, G., Nascimento, J., & Bradley, A. P. (2015). *Unregistered multiview mammogram analysis with pre-trained deep learning models.* Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Celebi, M. E., & Aslandogan, Y. A. (2005). *A comparative study of three moment-based shape descriptors.* Paper presented at the International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II.

Chan, H.-P., Sahiner, B., Petrick, N., Helvie, M. A., Lam, K. L., Adler, D. D., & Goodsitt, M. M. (1997). Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. *Physics in Medicine and Biology, 42*(3), 549.

Cheng, E., Xie, N., Ling, H., Bakic, P. R., Maidment, A. D., & Megalooikonomou, V. (2010). *Mammographic image classification using histogram intersection.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on.

Cheng, H.-D., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition, 36*(12), 2967-2991.

Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., & Du, H. (2006). Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition, 39*(4), 646-668.

Christoyianni, I., Dermatas, E., & Kokkinakis, G. (1999). *Neural classification of abnormal tissue in digital mammography using statistical features of the texture.* Paper presented at the Electronics, Circuits and Systems, 1999. Proceedings of ICECS'99. The 6th IEEE International Conference on.

Costa, D. D., Campos, L. F., Barros, A. K., & Silva, A. C. (2007). *Independent component analysis in breast tissues mammograms images classification using LDA and SVM.* Paper presented at the Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on.

Cruz-Roa, A., Díaz, G., Romero, E., & González, F. A. (2011). Automatic annotation of histopathological images using a latent topic model based on non-negative matrix factorization. *Journal of pathology informatics, 2*.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). *Visual categorization with bags of keypoints.* Paper presented at the Workshop on statistical learning in computer vision, ECCV.

D'Orsi, C. J. (2013). *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*.

da Rocha, S. V., Junior, G. B., Silva, A. C., de Paiva, A. C., & Gattass, M. (2016). Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution. *Expert Systems with Applications, 66*, 7-19.

Dai, L., Sun, X., Wu, F., & Yu, N. (2013). *Large scale image retrieval with visual groups.* Paper presented at the Image Processing (ICIP), 2013 20th IEEE International Conference on.

Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Dhungel, N., Carneiro, G., & Bradley, A. P. (2015). *Automated mass detection in mammograms using cascaded deep learning and random forests.* Paper presented at the Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on.

Diamant, I., Greenspan, H., & Goldberger, J. (2012). *Breast tissue classification in mammograms using visual words.* Paper presented at the Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of.

Doi, K. (2009). *Computer-aided diagnosis in medical imaging: achievements and challenges.* Paper presented at the World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany.

Doma, D. (2008). *Comparison of Different Image Interpolation Algorithms.* West Virginia University.

Drnasin, I., Gogić, G., & Drnasin, K. (2010). *Smartphone web radiology.* Paper presented at the Proceedings of CARS.

Drnasin, I., & Grgic, M. (2010). *The use of mobile phones in radiology.* Paper presented at the Proceedings of ELMAR.

Eakins, J. P., & Graham, M. E. (1999). Content-based image retrieval, a report to the JISC Technology Applications programme.

Eberl, M. M., Fox, C. H., Edge, S. B., Carter, C. A., & Mahoney, M. C. (2006). BI-RADS classification for management of abnormal mammograms. *The Journal of the American Board of Family Medicine, 19*(2), 161-164.

El-Faramawy, N., Rangayyan, R., Desautels, J., & Alim, O. (1996). *Shape factors for analysis of breast tumors in mammograms.* Paper presented at the Electrical and Computer Engineering, 1996. Canadian Conference on.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS One, 8*(5), e64366.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., & Equitz, W. (1994). Efficient and effective querying by image content. *Journal of intelligent information systems, 3*(3-4), 231-262.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International journal of cancer, 127*(12), 2893-2917.

Ferreira, C. B. R., & Borges, D. L. (2003). Analysis of mammogram classification using a wavelet transform decomposition. *Pattern Recognition Letters, 24*(7), 973-982.

Freer, T. W., & Ulissey, M. J. (2001). Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center 1. *Radiology, 220*(3), 781-786.

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer graphics and image processing, 4*(2), 172-179.

Gao, Q.-G., & Wong, A. (1993). Curve detection based on perceptual organization. *Pattern Recognition, 26*(7), 1039-1046.

Gonzalez, R., & Wintz, P. (1977). Digital image processing.

Görgel, P., Sertbas, A., & Uçan, O. N. (2015). Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines. *Expert Systems, 32*(1), 155-164.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

Halls, S. (2017, 17 April 2017). Mass shape, margin, and density as found with screening mammography. Retrieved from http://breast-cancer.ca/mass-chars/

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE, 67*(5), 786-804.

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*(6), 610-621.

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, W. P. (2000). *The digital database for screening mammography*. Paper presented at the Proceedings of the 5th international workshop on digital mammography.

Hologic Inc. ImageChecker.

Horsch, K., Giger, M. L., Vyborny, C. J., Lan, L., Mendelson, E. B., & Hendrick, R. E. (2006). Classification of Breast Lesions with Multimodality Computer-aided Diagnosis: Observer Study Results on an Independent Clinical Data Set 1. *Radiology, 240*(2), 357-368.

Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Schmidt, R. A., & Doi, K. (1998). Automated computerized classification of malignant and benign masses on digitized mammograms. *Academic Radiology, 5*(3), 155-168.

Hussain, M., Khan, S., Muhammad, G., Berbar, M., & Bebis, G. (2012). *Mass detection in digital mammograms using gabor filter bank.* Paper presented at the Image Processing (IPR 2012), IET Conference on.

Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis* (Vol. 46): John Wiley & Sons.

iCAD. (2000). Second Look. Retrieved from http://www.icadmed.com/secondlook-digital.html

Islam, M. J., Ahmadi, M., & Sid-Ahmed, M. A. (2010). An efficient automatic mass classification method in digitized mammograms using artificial neural network. *arXiv preprint arXiv:1007.5129*.

Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging, 37*(3), 420-426.

Jégou, H., Douze, M., & Schmid, C. (2010). Improving bag-of-features for large scale image search. *International journal of computer vision, 87*(3), 316-336.

Jelinek, H. F., Pires, R., Padilha, R., Goldenstein, S., Wainer, J., & Rocha, A. (2013). *Quality control and multi-lesion detection in automated retinopathy classification using a visual words dictionary.* Paper presented at the Intl. Conference of the IEEE Engineering in Medicine and Biology Society.

Jiang, M., Zhang, S., Li, H., & Metaxas, D. N. (2015). Computer-aided diagnosis of mammographic masses using scalable image retrieval. *Biomedical Engineering, IEEE Transactions on, 62*(2), 783-792.

Jirari, M. (2008). *Computer Aided System For Detecting Masses in Mammograms.* Kent State University.

Joseph, S., & Balakrishnan, K. (2011). Local binary patterns, haar wavelet features and haralick texture features for mammogram image classification using artificial neural networks *Advances in Computing and Information Technology* (pp. 107-114): Springer.

Juan, L., & Gwun, O. (2009). A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP), 3*(4), 143-152.

Julesz, B. (1975). Experiments in the visual perception of texture. *Scientific American, 232*, 34-43.

Jurie, F., & Triggs, B. (2005). *Creating efficient codebooks for visual recognition.* Paper presented at the Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on.

Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2007). Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools, 16*(06), 1047-1067.

Kelly, P. M., & Cannon, T. M. (1994). *CANDID: Comparison algorithm for navigating digital image databases.* Paper presented at the Scientific and Statistical Database Management, 1994. Proceedings., Seventh International Working Conference on.

Kelly, P. M., Cannon, T. M., & Hush, D. R. (1995). *Query by image example: the comparison algorithm for navigating digital image databases (CANDID) approach.* Paper presented at the IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology.

Khuzi, A. M., Besar, R., Zaki, W. W., & Ahmad, N. (2009). Identification of masses in digital mammogram using gray level co-occurrence matrices. *Biomedical imaging and intervention journal, 5*(3).

Kilday, J., Palmieri, F., & Fox, M. D. (1993). Classifying mammographic lesions using computerized image analysis. *IEEE Transactions on Medical Imaging, 12*(4), 664-669.

Kopans, D. B. (2007). *Breast imaging*: Lippincott Williams & Wilkins.

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 47-58.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Kulkarni, P., Stranieri, A., Kulkarni, S., Ugon, J., & Mittal, M. (Feb 2014). *Hybrid Technique based on N-gram and Neural Networks for Classification of Mammographic Images* Paper presented at the International Conference on Signal, Image Processing and Pattern Recognition, Sydney, Australia.

Kulkarni, P., Stranieri, A., Kulkarni, S., Ugon, J., & Mittal, M. (Mar 2014). Visual Character N-grams for Classification and Retrieval of Radiological Images. *The International Journal of Multimedia & Its Applications, 6*(2), 35.

Kulkarni, P., Stranieri, A., & Ugon, J. (2016). Pixel N-grams: Size, Location and Resolution Invariance for Shape Classification. *International Journal of Science, Engineering and Management, 1*(8), 38-44.

Kulkarni, P., Stranieri, A., Kulkarni, S., Ugon, J., & Mittal, M. (2015). *Analysis and Comparison of Co-occurrence Matrix and Pixel N-gram Features for Mammographic Images*. Paper presented at the International Conference on Communication and Computing, Banglore, India.

Kulkarni, P., Stranieri, A., Ugon, J. (Aug 2016). *Texture Image Classification using Pixel N-grams*. Paper presented at the IEEE International Conference on Signal and Image Processing, Beijing, China.

Kulkarni, P., Stranieri, A., Ugon, J., Kulkarni, S & Mittal, M. (April 2017). *Pixel N-grams for mammographic lesion classification*. Paper presented at the IEEE International Conference on Communication Systems, Computing and IT Applications, Mumbai, India.

LakeImaging. (2017). https://www.lakeimaging.com.au/.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(8), 1265-1278.

Lewenstein, K., & Urbaniak, K. (2016). Detection and evaluation of breast tumors on the basis of microcalcification analysis *Advanced Mechatronics Solutions* (pp. 159-166): Springer.

Li, H.-D., Kallergi, M., Clarke, L. P., Jain, V. K., & Clark, R. A. (1995). Markov random field for tumor detection in digital mammography. *IEEE Transactions on Medical Imaging, 14*(3), 565-576.

Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence, 25*(9), 1075-1088.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). *Composing simple image descriptions using web-scale n-grams.* Paper presented at the Proceedings of the Fifteenth Conference on Computational Natural Language Learning.

Li, T., Mei, T., Kweon, I.-S., & Hua, X.-S. (2011). Contextual bag-of-words for visual categorization. *Circuits and Systems for Video Technology, IEEE Transactions on, 21*(4), 381-392.

Li, Y., Chen, H., Rohde, G. K., Yao, C., & Cheng, L. (2015). Texton analysis for mass classification in mammograms. *Pattern Recognition Letters, 52*, 87-93.

Liu, J., Liu, X., Chen, J., & Tang, J. (2011). *Improved local binary patterns for classification of masses using mammography.* Paper presented at the Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on.

Liu, Y., Zhang, D., Lu, G., & Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262-282.

Lladó, X., Oliver, A., Freixenet, J., Martí, R., & Martí, J. (2009). A textural approach for mass false positive reduction in mammography. *Computerized Medical Imaging and Graphics, 33*(6), 415-422.

Løberg, M., Lousdal, M. L., Bretthauer, M., & Kalager, M. (2015). Benefits and harms of mammography screening. *Breast Cancer Research, 17*(1), 63.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Cruz-Roa, A., & González, F. A. (2013). *Bag-of-visual-ngrams for histopathology image classification.* Paper presented at the IX International Seminar on Medical Information Processing and Analysis.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision, 60*(2), 91-110.

Lu, S., & Bottema, M. J. (2003). *Structural image texture and early detection of breast cancer.* Paper presented at the Proceedings of the 2003 APRS Workshop on Digital Image Computing.

M-reader. (2001). MammoReader. Retrieved from https://sbir-cancercontrol.cancer.gov/sbir/viewProduct.do?prodId=896111

Marques, A. (1999). *Computerized classification of breast lesions: shape and texture analysis using an artificial neural network.* Paper presented at the Image Processing and Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465).

Martins, L. d. O., dos Santos, A. M., Silva, A. C., & Paiva, A. C. (2006). *Classification of normal, benign and malignant tissues using co-occurrence matrix and Bayesian neural network in mammographic images.* Paper presented at the 2006 Ninth Brazilian Symposium on Neural Networks (SBRN'06).

Materka, A., & Strzelecki, M. (1998). Texture analysis methods–a review. *Technical university of lodz, institute of electronics, COST B11 report, Brussels*, 9-11.

Mc Leod, P., & Verma, B. (2013). Variable Hidden Neuron Ensemble for Mass Classification in Digital Mammograms [Application Notes]. *IEEE Computational Intelligence Magazine, 8*(1), 68-76.

McKenzie, E. (2014). Breast Cancer Screening. Technical report

Metsälä, E., Pajukari, A., & Aro, A. R. (2012). Breast cancer worry in further examination of mammography screening–a systematic review. *Scandinavian journal of caring sciences, 26*(4), 773-786.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., . . . Van Gool, L. (2005). A comparison of affine region detectors. *International journal of computer vision, 65*(1-2), 43-72.

Mohamed, B. A., Salem, N. M., Hadhoud, M. M., & Seddik, A. F. (2016). Automatic Segmentation and Classification of Masses from Digital Mammograms. *Advances in Image and Video Processing, 4*(4), 17.

Mohanty, A. K., Champati, P. K., Swain, S. K., & Lenka, S. K. (2011). A review on computer aided mammography for breast cancer diagnosis and classification using image mining methodology. *International Journal of Computer Science and Communication, 2*(2), 531-538.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic Radiology, 19*(2), 236-248.

Mousa, R., Munib, Q., & Moussa, A. (2005). Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert Systems with Applications, 28*(4), 713-723.

Mu, T., Nandi, A. K., & Rangayyan, R. M. (2008). Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. *Journal of digital imaging, 21*(2), 153-169.

Mukanova, A., Hu, G., & Gao, Q. (2014). *N-Gram Based Image Representation and Classification Using Perceptual Shape Features.* Paper presented at the Computer and Robot Vision (CRV), 2014 Canadian Conference on.

Muramatsu, C., Li, Q., Suzuki, K., Schmidt, R. A., Shiraishi, J., Newstead, G. M., & Doi, K. (2005). Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results. *Medical physics, 32*(7), 2295-2304.

Muramatsu, C., Schmidt, R. A., Shiraishi, J., Li, Q., & Doi, K. (2010). Presentation of similar images as a reference for distinction between benign and malignant masses on mammograms: analysis of initial observer study. *Journal of digital imaging, 23*(5), 592-602.

Muramatsu, C., Zhang, M., Hara, T., Endo, T., & Fujita, H. (2014). *Differentiation of malignant and benign masses on mammograms using radial local ternary pattern.* Paper presented at the International Workshop on Digital Mammography.

Myatt, G. J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*: John Wiley & Sons.

Nakayama, R., Abe, H., Shiraishi, J., & Doi, K. (2009). Potential Usefulness of Similar Images in the Differential Diagnosis of Clustered Microcalcifications on Mammograms 1. *Radiology*, *253*(3), 625-631.

Nanni, L., Brahnam, S., & Lumini, A. (2012). A very high performing system to discriminate tissues in mammograms as benign and malignant. *Expert Systems with Applications, 39*(2), 1968-1971.

National Breast Cancer Foundation. (2016). Breast cancer stage 0 and stage 1,. *http://www.nationalbreastcancer.org/breast-cancer-stage-0-and-stage-1*.

Newton, E. V. (2016, 29-04-2016). Breast Cancer Screening *http://emedicine.medscape.com/article/1945498-overview#a5*.

Nister, D., & Stewenius, H. (2006). *Scalable recognition with a vocabulary tree.* Paper presented at the Computer vision and pattern recognition, 2006 IEEE computer society conference on.

Nithya, R., & Santhi, B. (2011). Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *International Journal of Computer Applications, 28*(6), 21-25.

Ojala, T., & Pietikäinen, M. (1999). Unsupervised texture segmentation using feature distributions. *Pattern Recognition, 32*(3), 477-486.

Ojansivu, V., & Heikkilä, J. (2008). *Blur insensitive texture classification using local phase quantization.* Paper presented at the International conference on image and signal processing.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica, 11*(285-296), 23-27.

Palazzetti, V., Guidi, F., Ottaviani, L., Valeri, G., Baldassarre, S., & Giuseppetti, G. (2016). Analysis of mammographic diagnostic errors in breast clinic. *La radiologia medica*, 1-6.

Papakostas, G. A., Boutalis, Y. S., Karras, D. A., & Mertzios, B. G. (2007). A new class of Zernike moments for computer vision applications. *Information Sciences, 177*(13), 2802-2819.

Pedrosa, G. V., Rahman, M. M., Antani, S. K., Demner-Fushman, D., Long, L. R., & Traina, A. J. (2014). *Integrating visual words as bunch of n-grams for effective biomedical image classification.* Paper presented at the Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on.

Pedrosa, G. V., & Traina, A. J. (2013). *From bag-of-visual-words to bag-of-visual-phrases using n-grams.* Paper presented at the Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on.

Pedrosa, G. V., Traina, A. J., & Traina, C. (2014). *Using sub-dictionaries for image representation based on the bag-of-visual-words approach.* Paper presented at the Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on.

Pelka, O., & Friedrich, C. M. (2015). FHDO biomedical computer science group at medical classification task of ImageCLEF 2015. *Working Notes of CLEF, 2015*.

Petrick, N., Chan, H. P., Sahiner, B., & Helvie, M. A. (1999). Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Medical physics, 26*(8), 1642-1654.

Petrosian, A., Chan, H.-P., Helvie, M. A., Goodsitt, M. M., & Adler, D. D. (1994). Computer-aided diagnosis in mammography: classification of mass and normal tissue by texture analysis. *Physics in Medicine and Biology, 39*(12), 2273.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). *Object retrieval with large vocabularies and fast spatial matching.* Paper presented at the Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.

Pow, R. E., Mello-Thoms, C., & Brennan, P. (2016). Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: A review of the evidence. *Journal of medical imaging and radiation oncology, 60*(3), 306-314.

Radiology, A. C. o. (2013). ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. *Reston, VA*.

Rahman, M. M., Antani, S. K., & Thoma, G. R. (2011). *Biomedical CBIR using "bag of keypoints" in a modified inverted index.* Paper presented at the Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on.

Rangayyan, R. M., Ayres, F. J., & Desautels, J. L. (2007). A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute, 344*(3), 312-348.

Rangayyan, R. M., El-Faramawy, N. M., Desautels, J. L., & Alim, O. A. (1997). Measures of acutance and shape for classification of breast tumors. *IEEE Transactions on Medical Imaging, 16*(6), 799-810.

Rangayyan, R. M., Mudigonda, N. R., & Desautels, J. L. (2000). Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical and Biological Engineering and Computing, 38*(5), 487-496.

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

Rickman, R., & Rosin, P. (1996). *Content-based image retrieval using colour n-grams.* Paper presented at the Intelligent Image Databases, IEE Colloquium on.

Rickman, R. M., & Stonham, T. J. (1996). *Content-based image retrieval using color tuple histograms.* Paper presented at the Electronic Imaging: Science & Technology.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Rocha, A., Carvalho, T., Jelinek, H. F., Goldenstein, S., & Wainer, J. (2012). Points of interest and visual dictionaries for automatic retinal lesion detection. *IEEE Transactions on Biomedical Engineering, 59*(8), 2244-2253.

Sampat, M. P., Bovik, A. C., & Markey, M. K. (2005). *Classification of mammographic lesions into BI-RADS shape categories using the beamlet transform.* Paper presented at the Medical Imaging.

Sampat, M. P., Markey, M. K., & Bovik, A. C. (2005). Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing, 2*(1), 1195-1217.

Scanis Inc. Mammex. Retrieved from https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapid=7928155

Selvakumar, K., & Ray, B. K. (2013). SURVEY ON POLYGONAL APPROXIMATION TECHNIQUES FOR DIGITAL PLANAR CURVES. *International Journal ofInfonnation Technology, Modeling and Computing (UITMC) Vol, 1*.

Setitra, I., & Larabi, S. (2015). *SIFT Descriptor for Binary Shape Discrimination, Classification and Matching.* Paper presented at the International Conference on Computer Analysis of Images and Patterns.

Sheshadri, H. S., & Kandaswamy, A. (2006). Breast tissue classification using statistical feature extraction of mammograms. *医用画像情報学会雑誌, 23*(3), 105-107.

Shyu, C.-R., Brodley, C., Kak, A., Kosaka, A., Aisen, A., & Broderick, L. (1998). *Local versus global features for content-based image retrieval.* Paper presented at the Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on.

Sickles, E. A., Miglioretti, D. L., Ballard-Barbash, R., Geller, B. M., Leung, J. W., Rosenberg, R. D., Yankaskas, B. C. (2005). Performance Benchmarks for Diagnostic Mammography 1. *Radiology, 235*(3), 775-790.

Sivic, & Zisserman. (2003). Video Google: a text retrieval approach to object matching in videos (pp. 1470-1477). USA.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Soltanian-Zadeh, H., & Rafiee-Rad, F. (2004). Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms. *Pattern Recognition, 37*(10), 1973-1986.

Spärck Jones, K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation, 60*(5), 521-523.

Stopel, D., Boger, Z., Moskovitch, R., Shahar, Y., & Elovici, Y. (2006). *Application of artificial neural networks techniques to computer worm detection.* Paper presented at the Neural Networks, 2006. IJCNN'06. International Joint Conference on.

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., . . . Kok, S. (1994). *The mammographic image analysis society digital mammogram database.* Paper presented at the Exerpta Medica. International Congress Series.

Suen, C. Y. (1979). N-gram statistics for natural language understanding and text processing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(2), 164-172.

Suleiman, W. I., Lewis, S. J., Georgian-Smith, D., Evanoff, M. G., & McEntee, M. F. (2014). Number of mammography cases read per year is a strong predictor of sensitivity. *Journal of Medical Imaging, 1*(1), 015503-015503.

Sun, Y., Todorovic, S., & Goodison, S. (2010). Local-learning-based feature selection for high-dimensional data analysis. *IEEE Transactions on pattern analysis and machine intelligence, 32*(9), 1610-1626.

Tan, X., & Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on, 19*(6), 1635-1650.

Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Transactions on Information Technology in Biomedicine, 13*(2), 236-251.

te Brake, G. M., Karssemeijer, N., & Hendriks, J. (1998). Automated detection of breast carcinomas not detected in a screening program. *Radiology, 207*(2), 465-471.

te Brake, G. M., Karssemeijer, N., & Hendriks, J. H. (2000). An automatic method to discriminate malignant masses from normal tissue in digital mammograms1. *Physics in Medicine and Biology, 45*(10), 2843.

Teuner, A., Pichler, O., & Hosticka, B. J. (1995). Unsupervised texture segmentation of images using tuned matched Gabor filters. *IEEE transactions on image processing, 4*(6), 863-870.

Tirilly, P., Claveau, V., & Gros, P. (2008). *Language modeling for bag-of-visual words image categorization.* Paper presented at the Proceedings of the 2008 international conference on Content-based image and video retrieval.

Tourassi, G. D., Harrawood, B., Singh, S., & Lo, J. Y. (2007). Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance. *Medical physics, 34*(8), 3193-3204.

Tourassi, G. D., Vargas-Voracek, R., Catarious Jr, D. M., & Floyd Jr, C. E. (2003). Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Medical physics, 30*(8), 2123-2130.

Tsai, C.-F. (2012). Bag-of-Words Representation in Image Annotation: A Review. *ISRN Artificial Intelligence, 2012*, 1-19. doi:10.5402/2012/376804

Uyun, S., Hartati, S., & Harjoko, A. (2013). Selection Mammogram Texture Descriptors Based on Statistics Properties Backpropagation Structure. *arXiv preprint arXiv:1307.6542*.

Vadivel, A., & Surendiran, B. (2013). A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories. *Computers in biology and medicine, 43*(4), 259-267.

van de Sande, K. E., Gevers, T., & Snoek, C. G. (2011). Empowering visual categorization with the GPU. *IEEE Transactions on Multimedia, 13*(1), 60-70.

Varela, C., Timp, S., & Karssemeijer, N. (2006). Use of border information in the classification of mammographic masses. *Physics in Medicine and Biology, 51*(2), 425.

Verma, B., McLeod, P., & Klevansky, A. (2009). A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. *Pattern Recognition, 42*(9), 1845-1852.

Wang, J., Li, Y., Zhang, Y., Xie, H., & Wang, C. (2011). *Bag-of-features based classification of breast parenchymal tissue in the mammogram via jointly selecting and weighting visual words.* Paper presented at the Image and Graphics (ICIG), 2011 Sixth International Conference on.

Wang, S., McKenna, M., Wei, Z., Liu, J., Liu, P., & Summers, R. M. (2013). Visual Phrase Learning and Its Application in Computed Tomographic Colonography *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013* (pp. 243-250): Springer.

Wang, Z., Yu, G., Kang, Y., Zhao, Y., & Qu, Q. (2014). Breast tumor detection in digital mammography based on extreme learning machine. *Neurocomputing, 128*, 175-184.

Wei, C.-H., Chen, S. Y., & Liu, X. (2012). Mammogram retrieval on similar mass lesions. *Computer methods and programs in biomedicine, 106*(3), 234-248.

Wei, C.-H., Li, C.-T., & Wilson, R. (2005). *A general framework for content-based medical image retrieval with its application to mammograms.* Paper presented at the Medical Imaging.

Wei, C.-H., Li, Y., & Huang, P. J. (2011). Mammogram retrieval through machine learning within BI-RADS standards. *Journal of biomedical informatics, 44*(4), 607-614.

Weka, W. (2011). 3: data mining software in Java. *University of Waikato, Hamilton, New Zealand (www. cs. waikato. ac. nz/ml/weka).*

Weszka, J. S., Dyer, C. R., & Rosenfeld, A. (1976). A comparative study of texture measures for terrain classification. *IEEE Transactions on Systems, Man, and Cybernetics*(4), 269-285.

Wong, M. T., He, X., Nguyen, H., & Yeh, W.-C. (2012). *Mass classification in digitized mammograms using texture features and artificial neural network.* Paper presented at the International Conference on Neural Information Processing.

World Cancer Research Fund International. (2017). Breast cancer statistics. Retrieved from http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics

World Health Organisation. (2013). Latest world cancer statistics, International Agency for Research on Cancer (IARC) press releases 2013, .

Xinlily, W., & Bent, O. (1994). Texture features from gray level gap length matrix.

Yadav, R. B., Nishchal, N. K., Gupta, A. K., & Rastogi, V. K. (2007). Retrieval and classification of shape-based objects using Fourier, generic Fourier, and wavelet-Fourier descriptors technique: A comparative study. *Optics and Lasers in engineering, 45*(6), 695-708.

Yang, M., Kpalma, K., & Ronsin, J. (2008). A survey of shape feature extraction techniques. *Pattern Recognition*, 43-90.

Yang, W., Lu, Z., Yu, M., Huang, M., Feng, Q., & Chen, W. (2012). Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single-and multiphase contrast-enhanced CT images. *Journal of digital imaging, 25*(6), 708-719.

You, Z., & Jain, A. K. (1984). Performance evaluation of shape matching via chord length distribution. *Computer vision, graphics, and image processing, 28*(2), 185-198.

Zhang, D., & Lu, G. (2001). *A comparative study on shape retrieval using Fourier descriptors with different shape signatures.* Paper presented at the Proc. International Conference on Intelligent Multimedia and Distance Education (ICIMADE01).

Zhang, J., & Tan, T. (2002). Brief review of invariant texture analysis methods. *Pattern Recognition, 35*(3), 735-747.

Zhang, Y., Tomuro, N., Furst, J., & Raicu, D. S. (2012). Building an ensemble system for diagnosing masses in mammograms. *International journal of computer assisted radiology and surgery, 7*(2), 323-329.

Zheng, B. (2009). Computer-aided diagnosis in mammography using content-based image retrieval approaches: current status and future perspectives. *Algorithms, 2*(2), 828-849.

Zheng, Q.-F., Wang, W.-Q., & Gao, W. (2006). *Effective and efficient object-based image retrieval using visual phrases.* Paper presented at the Proceedings of the 14th annual ACM international conference on Multimedia.

Zhu, L., Rao, A., & Zhang, A. (2002). Advanced feature extraction for keyblock-based image retrieval. *Information Systems, 27*(8), 537-557.