# ADAPTIVE FEATURE ENGINEERING MODELING FOR ULTRASOUND IMAGE CLASSIFICATION FOR DECISION SUPPORT

by

Hatwib Mugasa, B.Sc., M.S.

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

COLLEGE OF ENGINEERING AND SCIENCE
LOUISIANA TECH UNIVERSITY

November 2019

LOUISIANA TECH UNIVERSITY

**GRADUATE SCHOOL**

**September 20, 2019**
Date of dissertation defense

We hereby recommend that the dissertation prepared by

**Hatwib Mugasa, B.Sc, M.S.**

entitled **ADAPTIVE FEATURE ENGINEERING MODELING FOR ULTRASOUND IMAGE CLASSIFICATION FOR DECISION SUPPORT**

be accepted in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy in Computational Analysis & Modeling**

Dr. Sumeet Dua, Supervisor of Dissertation Research

Dr. Weizhong Dai,
Head of Computational Analysis & Modeling

**Members of the Doctoral Committee:**
Dr. Sumeet Dua
Dr. Weizhong Dia
Dr. Pradeep Chiorwappa
Dr. Jinko Kanno
Dr. Box Leangsuksun

**Approved:**

Hisham Hegab
Dean of Engineering & Science

**Approved**:

Ramu Ramachandran
Dean of the Graduate School

GS Form 13a
(10/18)

# ABSTRACT

Ultrasonography is considered a relatively safe option for the diagnosis of benign and malignant cancer lesions due to the low-energy sound waves used. However, the visual interpretation of the ultrasound images is time-consuming and usually has high false alerts due to speckle noise. Improved methods of collection image-based data have been proposed to reduce noise in the images; however, this has proved not to solve the problem due to the complex nature of images and the exponential growth of biomedical datasets. Secondly, the target class in real-world biomedical datasets, that is the focus of interest of a biopsy, is usually significantly underrepresented compared to the non-target class. This makes it difficult to train standard classification models like Support Vector Machine (SVM), Decision Trees, and Nearest Neighbor techniques on biomedical datasets because they assume an equal class distribution or an equal misclassification cost. Resampling techniques by either oversampling the minority class or under-sampling the majority class have been proposed to mitigate the class imbalance problem but with minimal success. We propose a method of resolving the class imbalance problem with the design of a novel data-adaptive feature engineering model for extracting, selecting, and transforming textural features into a feature space that is inherently relevant to the application domain.

We hypothesize that by maximizing the variance and preserving as much variability in well-engineered features prior to applying a classifier model will boost the differentiation of the thyroid nodules (benign or malignant) through effective model building. Our proposed a hybrid approach of applying Regression and Rule-Based techniques to build our Feature Engineering and a Bayesian Classifier respectively.

In the Feature Engineering model, we transformed images pixel intensity values into a high dimensional structured dataset and fitting a regression analysis model to estimate relevant kernel parameters to be applied to the proposed filter method. We adopted an Elastic Net Regularization path to control the maximum log-likelihood estimation of the Regression model. Finally, we applied a Bayesian network inference to estimate a subset for the textural features with a significant conditional dependency in the classification of the thyroid lesion. This is performed to establish the conditional influence on the textural feature to the random factors generated through our feature engineering model and to evaluate the success criterion of our approach.

The proposed approach was tested and evaluated on a public dataset obtained from thyroid cancer ultrasound diagnostic data. The analyses of the results showed that the classification performance had a significant improvement overall for accuracy and area under the curve when then proposed feature engineering model was applied to the data. We show that a high performance of 96.00% accuracy with a sensitivity and specificity of 99.64%) and 90.23% respectively was achieved for a filter size of $13 \times 13$.

# APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Thesis. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Thesis. Further, any portions of the Thesis used in books, papers, and other works must be appropriately referenced to this Thesis.

Finally, the author of this Thesis reserves the right to publish freely, in the literature, at any time, any or all portions of this Thesis.

Author _____

Date _____

# DEDICATION

To the ones who uphold community, knowledge, and family interests with great love, care, kindness, and honesty.

# TABLE OF CONTENTS

ABSTRACT ..... iii

DEDICATION ..... vi

LIST OF FIGURES ..... xi

LIST OF TABLES ..... xiii

ACKNOWLEDGMENTS ..... xiv

CHAPTER 1 INTRODUCTION ..... 1

1.1 Thyroid Cancer ..... 1

1.2 Biomedical Informatics (BMI) ..... 2

1.2.1 Biomedical Imaging Techniques ..... 3

1.2.2 Clinical Decision Support Systems (CDSS) ..... 3

1.3 Diagnostic Imaging Techniques ..... 4

1.3.1 Digital-Based Imaging Techniques ..... 5

1.3.2 Ultrasound (US) Imaging ..... 6

1.3.3 Imaging of Thyroid Nodules ..... 7

1.4 Problem Statement ..... 8

1.4.1 Goals and Overview ..... 11

CHAPTER 2 RELATED WORKS ..... 14

2.1 Image Thresholding ..... 14

2.1.1 Challenges in Image Thresholding ..... 15

2.2 Image Filtering ..... 16
2.2.1 Correlation and Convolution ..... 16
2.2.2 Image Blurring ..... 17
2.2.3 Finding Contours ..... 17
2.2.4 Edge Detection ..... 18
2.2.5 Discrete Wavelet Transform (DWT) ..... 18
2.2.6 Histogram-Based Method ..... 19
2.3 Feature Extraction ..... 20
2.3.1 Binarized Statistical Image Features (BSIF) ..... 21
2.3.2 Haralick Textural Features ..... 22
2.4 Image Classifiers ..... 22
2.4.1 Logistic Based Classifiers ..... 26
2.4.2 Rule-Based Classifiers ..... 27
2.4.2.1 Bayesian Classifiers ..... 28
CHAPTER 3 RESEARCH METHODOLOGY ..... 30
3.1 Research Approach ..... 32
3.2 Data Preparation ..... 32
3.3 Image Pre-processing ..... 33
3.4 Proposed Image Filter Design ..... 34
3.4.1 Regression Model ..... 34
3.4.1.1 Linear Regression ..... 34
3.4.1.2 Singular Value Decomposition (SVD) ..... 35
3.4.1.3 Logistic Regression ..... 36

3.4.2 Regularization ..... 38

3.4.2.1 Ridge Regularization ..... 38

3.4.2.2 Least Absolute Shrinkage and Selection Operator Regularization(LASSO) ..... 38

3.4.2.3 Elastic Net Regularization ..... 39

3.4.2.4 Coefficient Shrink Factor ..... 39

3.4.3 Estimating Kernel Parameters ..... 39

3.4.4 Spectral Analysis ..... 40

3.5 Feature Engineering ..... 42

3.5.1 Feature Extraction ..... 42

3.5.1.1 Haralick Texture Features ..... 42

3.5.2 Feature Selection ..... 43

3.5.2.1 Feature Selection with Principal Component Analysis (PCA) ..... 43

3.5.2.2 Boruta Feature Selection Algorithm ..... 44

CHAPTER 4 BAYESIAN BASED CLASSIFICATION ..... 46

4.1 Bayesian Networks (BN) ..... 46

4.2 Markov Chains ..... 48

4.3 Bayesian Classifiers ..... 50

4.3.0.3 Parameter Learning ..... 52

4.3.0.4 Structure Learning ..... 52

4.4 Bayesian Model ..... 53

CHAPTER 5 EXPERIMENTAL RESULTS ..... 56

5.1 Experimental Data ..... 56

5.2 Image Filter ..... 60

5.3 Feature Selection ..... 60

5.3.1 Kruskal-Wallis Test for Feature Selection ..... 60

5.3.2 Proposed Bayesian Network Structure (BNS) ..... 63

5.4 Model Validation ..... 67

5.4.1 Classifier Evaluation ..... 68

5.4.2 Entropy Uncertainty ..... 68

5.5 Results ..... 69

5.6 Discussion ..... 70

CHAPTER 6 CONCLUSIONS ..... 77

6.1 Conclusions ..... 77

6.1.1 Model Complexity ..... 78

6.2 Future Work ..... 79

APPENDIX A REGULARIZATION ..... 81

APPENDIX B CROSS VALIDATION ..... 83

APPENDIX C IMAGE FILTERING WITH THE DISCRETE FOURIER TRANSFORM (DFT) ALGORITHM ..... 85

APPENDIX D FAST FOURIER TRANSFORM ..... 87

APPENDIX E CONVOLUTION IMAGE FILTER ALGORITHM ..... 89

APPENDIX F ALGORITHM FOR EXTRACTING HARALICK IMAGE FEATURES ..... 91

APPENDIX G BAYESIAN APPROACH ..... 93

LIST OF ABBREVIATIONS ..... 96

BIBLIOGRAPHY ..... 97

# LIST OF FIGURES


Figure 1.1: Evidence-based medicine..... 2

Figure 1.2: Evidence-based evaluation and treatment of patient thyroid nodules [8]... 9

Figure 1.3: A typical image of a healthy and cancerous thyroid..... 11

Figure 1.4: Ultrasound images of (a) benign and (b) malignant thyroid lesions with the Regions of Interest (ROI) marked in red..... 12

Figure 3.1: Block diagram for the proposed Computer-Aided Diagnosis (CAD) System..... 31

Figure 3.2: A Multilevel dimensional dataset using 3 Binarized Statistical Image Features (BSIF) filter levels..... 33

Figure 3.3: Data matrix in the BSIF features space..... 33

Figure 3.4: Transformation of image patches to a 1-D data frame..... 40

Figure 3.5: 2-D gray-scale intensity of ultrasound image after a fourier transformation. 41

Figure 3.6: System architecture diagram of the Boruta feature selection algorithm..... 44

Figure 4.1: Illustration of a joint probability distribution for events: A, B, C, D and E 48

Figure 4.2: Graph representation of a 3-state Markov Chain..... 50

Figure 4.3: A Bayesian inference network illustrating the conditional probability of the weather forecast..... 51

Figure 4.4: Posterior distributions means of the proposed model using Monte Carlo Markov Chain..... 55

Figure 4.5: Model performance diagnostic with the Hamiltonian energy divergences. 55

Figure 5.1: A screen shot of the computer imaging and medical applications laboratory system[62]..... 57

Figure 5.2: Ultrasound images of benign thyroid lesions..... 58

Figure 5.3: Expert annotations of benign Figure 5.3.a and malignant Figure 5.3.b thyroid lesions [62]. ..... 59

Figure 5.4: Ultrasound images of benign thyroid lesions..... 59

Figure 5.5: Ultrasound images of malignant thyroid lesions [62]..... 60

Figure 5.6: Target region with sliding filter window. ..... 61

Figure 5.7: Transformation of image pixel data into a structured data set..... 61

Figure 5.8: Illustration of kernel estimates and their respective filter transformation .. 62

Figure 5.9: Mean features Kruskal measure of statistical significant..... 63

Figure 5.10: Principal Component Importance levels for all features..... 64

Figure 5.11: Bayesian network with Spatial-independent feature properties..... 64

Figure 5.12: Bayesian network with Shape feature properties..... 65

Figure 5.13: Bayesian network of Moment feature properties..... 66

Figure 5.14: Filter performances with the Random Forest and Recursive Partitioning classifiers..... 70

Figure 5.15: Average model performance analysis plot..... 71

Figure 5.16: Model performance of Recursive Partitioning compared to the Random Forest classifier ..... 76

Figure 6.1: The relationship between complexity of different classes..... 79

# LIST OF TABLES


Table 1.1: Evolutionary history of diagnostic imaging techniques since the last 50 years ..... 7

Table 1.2: Characteristics of ultrasound thyroid nodule features [8]. ..... 10

Table 2.1: Haralick texture features notations.[23] ..... 23

Table 2.2: Haralick texture features notations (continued).[23] ..... 24

Table 2.3: Gray-level co-occurrence matrix (GLCM) notation[23]. ..... 25

Table 2.4: Gray-level co-occurrence matrix (GLCM) notation (continued)[23]. ..... 26

Table 5.1: The Thyroid Imaging Reporting and Data System (TI-RADS) scale used to describe thyroid lesions ..... 58

Table 5.2: Evaluation of feature preprocessing techniques fitted to classification models ..... 65

Table 5.3: Evaluation of the proposed filter method fitted to various classifiers. ..... 70

Table 5.4: Evaluation of the proposed filter method fitted to various classifiers (continued). ..... 72

Table 5.5: Selected studies on the CAD system for automated diagnosis of thyroid lesions (benign and malignant) with ultrasound images. ..... 73

Table 5.6: Selected studies on the CAD system for automated diagnosis of thyroid lesions (benign and malignant) with ultrasound images (continued). ..... 74

Table G-1: Table of symbols used in this study. ..... 94

Table G-2: Bayesian Theory notations ..... 95

# ACKNOWLEDGMENTS

It is with great pleasure to earn my doctorate degree at Louisiana Tech University. The completion of this undertaking would not be complete without the support of family and friends. I owe a deep sense of gratitude to my sister and her family for the unconditional support and encouragement with my studies.

I would like to express my heartfelt gratitude to Dr. Sumeet Dua, who introduced me to research and the scientific method. It is with great privilege to have studied under him; I will always treasure his advice and expertise. I would like to give special thanks to my committee members Dr. Weizhong Dai, Dr. Pradeep Chowriappa, Dr. Jinko Kanno, and Dr. Box Leangsuksun and all the faculty at Louisiana Tech University for their wonderful courses that helped me in my studies.

Last, but definitely not the least, I am deeply thankful to all the students and my friends who supported me, especially Ayesha Akter, Radhika Medury, Richard Appiah, Ali Alqahtani, Norman Mapes, and to Louisiana Tech University staff, especially Dr. Collin Wick, Dan Erickson, Marsha Smith, Natalie Osborne, and Bill Jones.

# CHAPTER 1

# INTRODUCTION

Information Technology (IT) is considered to be an integral part of the workflow process in many disciplines and has inspired rapid growth in various application domains such as Health Information Technology (Health-IT) and Biomedical Informatics (BMI). Health-IT refers to IT technology and infrastructure used for the design, storage, exchange, and analysis of electronic health information. Methods in Health-IT have also been successfully applied in Biomedical Informatics to understand the fundamentals of human biology and unravel the complexities in the diagnosis, treatment, and prevention of diseases, as illustrated in Figure 1.1.

## 1.1 Thyroid Cancer

The thyroid gland is a butterfly-shaped organ in the endocrine system responsible the production and secretion of thyroid hormones, namely, thyroxine (T4), and triiodothyronine (T3) hormones [2]. The main function of thyroid hormones is to regulate the body's metabolic activity and ensure the normal growth and development of other organs like the brain, heart, liver and muscles, particularly in infants. There levels in the bloodstream are regulated by the both hypothalamus and pituitary gland through the release of thyroid-stimulating hormone (TSH). Thus, a drop in T3 & T4 hormone blood levels will trigger an

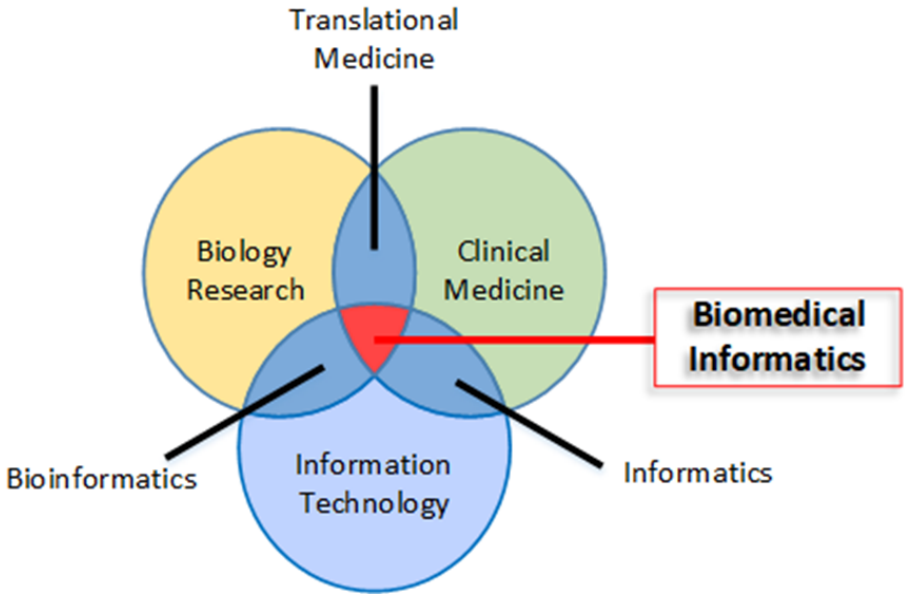



**Figure 1.1** Ecosystem of Evidence-based medicine [1]

increase in the production of TSH hormone, which in turn will stimulate the thyroid gland to produce more T3 & T4 hormones. In contrast, a rise in T3 & T4 hormone blood levels triggers the production of TSH hormone to halt [3], however, an imbalance could occur if the thyroid gland is unable to regulate T3 & T4 hormone production.

Studies show that thyroid hormone imbalance could lead to potential life-threatening health conditions such as hypothyroidism (under-active thyroid) or hyperthyroidism (over-active thyroid) [3]. The most common cause of hyperthyroidism and hypothyroidism is abnormal growth of thyroid cells, forming nodules within the thyroid gland.

## 1.2 Biomedical Informatics (BMI)

BMI is an interdisciplinary science encompassing the deployment of Health Information Technology (Health-IT) alongside Biological knowledge-bases and human interaction. It applies computational techniques on biomedical data for the purpose of

understanding and solving biological problems [4], with emphasis on data, information, and knowledge.

The current progress in BMI research is been driven by the significant improvement in technologies for management and analysis of data, leading to an exponential growth in the volume of healthcare data and development in biomedical imaging and processing techniques.

### 1.2.1 Biomedical Imaging Techniques

Biomedical images are a visual representation of the internal structure of a body like muscles, organ tissues, and bones. Images in clinical datasets are represented as usually as 2 or 3-dimensional arrays pixels of varying texture and color. The datasets go through well-designed data collection procedures by specialized physician. It is common to have abnormal growth of cell tissue leading to the formation of nodules. A biopsy is performed to confirm whether a nodule is benign or malignant. To assist in the diagnosis process, Clinical Decision Support Systems (CDSS) are used to improve the time taken and minimize the misdiagnosis of non-malignant nodules.

### 1.2.2 Clinical Decision Support Systems (CDSS)

A CDSS is any computer system whose aim is to provide knowledge to health care providers in support of their decision to maintain or improve the health of human beings. The knowledge obtained from CDSS is used to support the decision-making process for health care providers when diagnosing and/or treating an illness or physical injury of patients. It is also beneficial in determining the type of clinical data and diagnosis procedures required to perform specific health care test. In addition, CDSS can be used

as a guide by senior management to determine health-related risk and financial impact of decisions made by clinicians.

One of the early adoptions of a CDSS by Staniland et al. [5] was used to study the surgical and laboratory diagnostic process performed at the University of Leeds to explain the seven possible causes of acute abdominal pain. They used Bayesian Probability Theory and assumed conditional independence of seven causes of sudden offset of acute pain for various diagnosis procedures and had a diagnosis accuracy of 91.8% compared to 65%-80% by the clinicians.

Tools used in CDSS are classified in five categorizes according to the functions that affect their implementation, i.e.:

- their intended use,
- the process in which advice is provided,
- the consultations style by physicians,
- the decision-making process applied, and
- the level of human-computer interaction.

## 1.3 Diagnostic Imaging Techniques

In 1895, while studying the effect of a high voltage current passed through a charged cathode tube, Willhelm Röntgen discovered that a new kind of invisible ray would pass through an opaque black sheet of paper that was wrapped around the tube. In later experiments, he discovered that the invisible ray of light, known as an x-ray, could be used to expose the structure of bones and metals when passed through human tissue. In this

section, we briefly describe the history of diagnostic imaging techniques and illustrate its evolution in Table 1.1.

### 1.3.1 Digital-Based Imaging Techniques

In Radiography, an image pattern is produced when a beam of electrons with varying amplitude is released from an x-ray's tube through a target area. This variation in amplitude is responsible for producing the image when the electrons are captured on a medium or screen film. Fluoroscopy is a variant of the radiography imaging technique that projects x-ray beams at a lower frequency rate than the traditional radiography technique and exposes an image on a fluorescent screen. Despite the significant progress of clinical diagnosis brought about by the introduction of Radiographic image and low-quality fluoroscopic images, their 30% rate of misidentifying a lesion was considered high and insufficient.

The medical imaging has significant improvement from the time of x-rays to producing real-time 3-D images of the human body. Digital-based imaging techniques like the Computed Tomography (CT), Magnetic Resonance Image (MRI) and Ultrasound (US), were later introduced to produce near-real-time results and resolve issues of high misdiagnosis rates experienced with photographic film-based techniques.

Unlike Radiography, the x-rays in CT are produced by a computer detector system that eliminates electronic noise. The single x-ray detector can be used to produce a flat view of a target image by setting it in a fixed position. Multiple rotate detector can also be used to produce a 3-D view of the target image. In the MRI technique, an illustration of the chemical composition of a body tissue is produced by applying an oscillating magnetic

field over a weak radio frequency signal whereby producing hydrogen atoms in the body. An image is then generated by determining the variation of the magnetic field and rate at which the atoms return to an equilibrium. MRI is considered relatively safer than x-ray based techniques because of its non-ionizing properties.

### 1.3.2 Ultrasound (US) Imaging

US Imaging technique is based on sound waves to produce the internal composition of a human body. It works by exposing a high frequency sound wave of 1 to 5 MHz through the target body region and reflected the waves back to the emitting device. The distance and time traveled by the sound over alternating regions of low and high pressure are used to determine the density and structure of the image.

Despite the fact that US diagnostic imaging is considered to be the safest and least convoluted, it is prone to common image degradation errors like acoustic and speckle patterns caused by echo interference [7]. Secondly, in a real-world biomedical dataset, the target object that is the focus of interest of a biopsy is usually significantly underrepresented compared to the non-target object and artifacts in the image. This makes it difficult to train traditional classification models on biomedical datasets because they assume an equal class distribution or an equal misclassification cost.

The low probability of a specific cancerous tumor to be observed in the general population is also another cause of class imbalance in biomedical datasets. Various feature extraction and resampling techniques have been proposed to mitigate the class imbalance problem by either oversampling the minority class or under-sampling majority class;

**Table 1.1** Evolutionary history of diagnostic imaging techniques since last 50 years [6]

| | |
|---|---|
| 1950-1960 | Fluoroscopic ImageIntensifier and<br>Gamma Camera for Radionuclide Imaging |
| 1960-1970 | Automated Film Processor |
| 1970-1980 | Advanced Projection Radiography,<br>Digital Subtraction Angiography (DSA),<br>CT<br>US |
| 1980-1990 | Computed Radiography (CR),<br>MRI,<br>Color Doppler Ultrasound |
| 1990-2000 | Flat Panel Detector (FPD) Systems,<br>Magnetic Resonance Angiography (MRA),<br>Ultrafast-MRI,<br>Positron Emission Tomography (PET) and<br>Tissue Contrast Harmonic Imaging (CHI) |
| 2000-todate | Realtime 3D Ultrasound Imaging,<br>Parallel MRI, PET\CT,<br>Molecular Imaging, and<br>Picture Archiving and Communication Systems (PACS) |

however, this could lead to the generation of a high-dimension dataset, and a problem in selecting an appropriate resampling technique that maintains feature relevance.

### 1.3.3 Imaging of Thyroid Nodules

The American Association of Clinical Endocrinologists (AACE), American College of Endocrinology (ACE) and the Associazione Medici Endocrinologi (AME) [8] designed a set of standardized guidelines for identifying relevant anatomical characteristics thyroid nodule features from extensive literature reviews and physician's expert opinion.

The motivation for the guidelines is to optimize the clinical diagnosis process of thyroid lesions by ranking the risk of malignancy into three categories, i.e., benign, indeterminate, or suspicious for malignancy (see Table 1.2). The guidelines provide recommendations to applied by clinical practitioners, such as:

- confirming the presence of a thyroid nodule when the physical evaluation is ambiguous;
- evaluation and identification of thyroid nodule internal structures;
- rating the malignancy risk of thyroid nodules based on their appearance;
- radioisotope scanning on patients with low thyroid-stimulating hormone levels or with patients from geographic regions with iodine deficiency; and
- to provide recommendations prior to performing a surgical treatment.

## 1.4 Problem Statement

It is common to have abnormal growth of thyroid cells in the tissues of the thyroid gland which introduces the formation of nodules in the thyroid [2]. Approximately 90% of the population is likely to have at least one thyroid nodule by the time we reach 80 years old. The probability of developing a thyroid nodule regardless of be benign (non-cancerous) or malignant (cancerous) is higher as we age [9]. It is also noted that most of the thyroid nodules are benign while only a small percentage are malignant.

Benign thyroid nodules are non-life-threatening hence treatment is not necessary unless the nodule grows too big that it is difficult to breathe. Nonetheless, it is crucial to detect the malignant thyroid nodules at an early stage and to seek immediate treatment as shown in Figure 1.2 [2]. Generally, there are four main types of thyroid cancers. They are

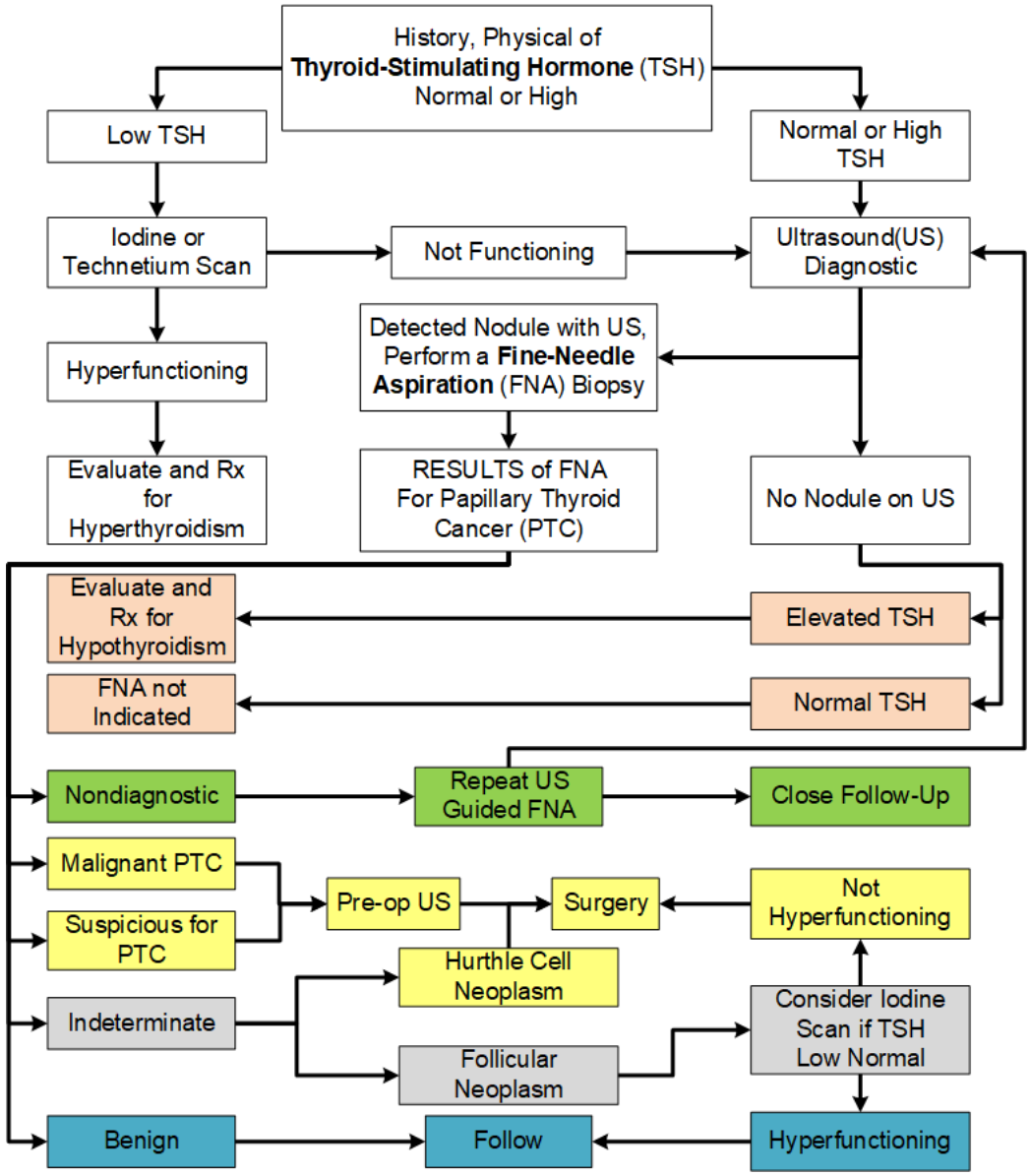



**Figure 1.2** Evidence-based evaluation and treatment of patient thyroid nodules [8].

namely, anaplastic, follicular, medullary, and papillary thyroid cancer [2]. These four types of cancers are based on characteristics of the cancer cells when viewed under a microscope.

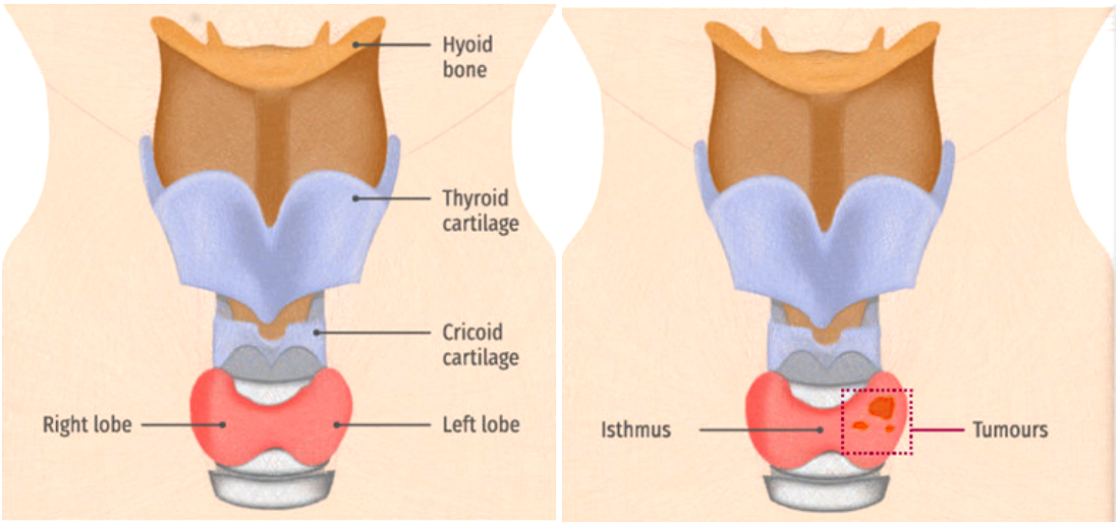
However, the growth of these thyroid nodules exhibits almost no symptoms [9]. Therefore, they can only be discovered during a routine health examination. It would probably be too late for treatment if the malignant thyroid nodules are detected at an advanced stage. According to the American Cancer Society [10], it was estimated that

**Table 1.2** Characteristics of ultrasound thyroid nodule features [8].

| **Benign Features** |
|---|
| Isoechoic spongiform appearance (microcystic spaces comprising of >50% of the nodule) |
| Simple cyst with thin regular margins |
| Mostly cystic (>50%) nodules containing colloid (hyperechoic spots with a comet-tail sign) |
| Regular "Eggshell" calcification around the periphery of a nodule. |
| **Malignant Features** |
| Papillary Carcinoma:<br>– Solid hypoechoic (relative to prethyroid muscles) nodule, which may contain hyperechoic foci without posterior shadowing (i.e., microcalcifications);<br>– Solid hypoechoic nodule, with intranodular vascularity and absence of peripheral halo;<br>– "Taller-than-wide" nodule (anteroposterior >transverse diameter when imaged in the transverse plane). |
| Hypoechoic nodule with spiculated or lobulated margin |
| Hypoechoic mass with a broken calcified rim and tissue extension beyond the calcified margin |
| Follicular neoplasm (either follicular adenoma or carcinoma) |
| Isoechoic or mildly hypoechoic homogeneous nodule with intranodular vascularization and well-defined halo |
| **Indeterminate Features** |
| Isoechoic or hyperechoic nodule with hypoechoic halo |
| Mild hypoechoic (relative to surrounding parenchyma) nodule with smooth margin |
| Peripheral vascularization |
| Intranodular macrocalcification |

there were 56,870 new cases of thyroid cases in the United States in the year 2017. It was also reported that the number of deaths due to thyroid cancer that year was 2,010.

The images in Figure 1.3 are visual representations of a healthy thyroid (shown in Figure 1.3.a) and tumors in a cancerous thyroid (shown in Figure 1.3.a). It can be clearly seen that a cancerous thyroid has nodules growing in the image for thyroid cancer. Whereas there is no abnormal growth of nodules in the image with a healthy thyroid, it is noted that not all the nodules formed are cancerous. A major challenge in classification of thyroid US



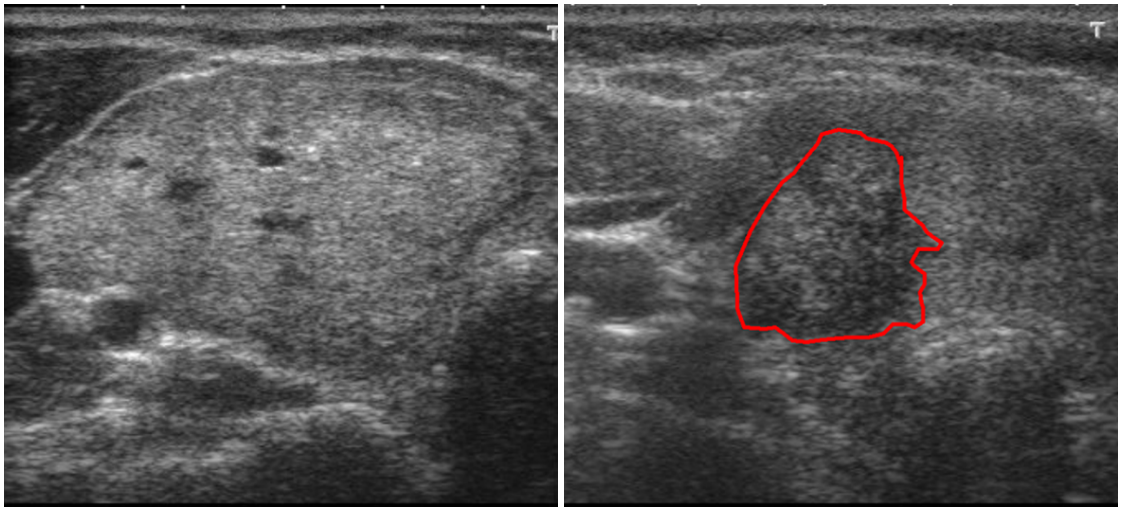

**(a)** A healthy thyroid **(b)** Cancerous thyroid

**Figure 1.3** A typical image of a healthy Figure 1.3.a and cancerous Figure 1.3.b thyroid

images is the degree of correctness due to the high variation in the follicular composition of thyroid nodules and low quality of US image. This poses a risk of a high rate of false-positive diagnosis leading to unnecessary Fine-Needle Aspiration (FNA) biopsy and/or medical over-treatment. A typical benign and malignant thyroid lesion image is represented in Figure 1.4.

### 1.4.1 Goals and Overview

Ultrasound equipment are common medical imaging modality tools used to locate and evaluate thyroid lesions [11] using Ultrasonography. It is a safe, non-ionizing, and



**(a)** Benign follicular lesion **(b)** Malignant

**Figure 1.4** Ultrasound images of (a) benign and (b) malignant thyroid lesions with the Regions of Interest (ROI) marked in red

non-invasive technique to examine the thyroid glands [12]. However, the medical diagnosis images captured with the ultrasound machine are indistinct and are not easily differentiable, especially for benign and malignant thyroid lesions (see Figure 1.4). A study of data management techniques for CDSS [13] shows that the most common factors that influence the quality of data in medical imaging are: (i) accuracy or correctness (ii) completeness (iii) timeliness and, (iv) the quality of evidence. Thus, the visual inspection of the ultrasound images must be analyzed by ultrasonographers who have adequate experience and expertise to ensure an accurate diagnosis. Otherwise, the diagnosis of the ultrasound images may be inaccurate and subjective. We propose to develop a computer-aided diagnosis (CAD) system that will mitigate this challenge in a twofold objective, namely:

- to assist ultrasonographers in giving an objective second-opinion of the thyroid lesion diagnosis (to determine if it is benign or malignant);

- and to make the proposed CAD system time-efficient by minimizing any manual interpretation of the images that is laborious.

In Chapter 2, we describe the background in the classification of biomedical informatics image data and the choice of our research approach. In Chapter 3, we describe an adaptive feature engineering model for classification of ultrasound diagnostic images. We hypothesize that minimizing the variance and maximizing correlation among textural features addresses this classification problem. In Chapter 4, we present an image classifier model using Bayesian inference based on the knowledge discovered from artifacts of the ultrasound images. In Chapter 5 we demonstrate the usefulness of automatically learning the structure of Bayesian networks by solving a classification problem from clinical decision support, namely, the classification of Ultrasound Images for Clinical Decision Support. We evaluate our model on a public dataset obtained from ovarian and thyroid cancer ultrasound diagnostic data in the final chapter.

# CHAPTER 2

# RELATED WORKS

The textural features of the diagnostic images were extracted using the following approaches from related literature. Textural features are useful in capturing morphological features that are reflected as nonlinear changes in texture in images. These nonlinear changes are captured as suitable texture features that can quantify the changes in the intensity, regularity, coarseness, contrast, homogeneity, etc. of the pixels of the image. In this chapter, we review popular extraction techniques used for extracting texture descriptors, from the simplest techniques like thresholding to complex rule-based methods for extracting textural characteristics of an image. We later present classification techniques that have been proposed in CAD systems for image processing of medical image data.

## 2.1 Image Thresholding

Thresholding is a process of creating a binary image from a grayscale one.[14] It is the simplest form of segmentation (separating an image into regions). It uses the simplest property that pixels in a region can share the intensity. Hence, thresholding is a natural way of separating light and dark regions of the image. In simple words, all pixels with intensity value below some threshold are being assigned zero and all pixels with intensity above this

threshold one. Pixels that are equal the threshold value are treated either as zero or one but the behavior needs to be consistent.

Let $f(x, y)$ be the gray level of a pixel $(x, y)$ and $T$ be the threshold value, then the thresholded image $g(x, y)$ is defined as:

$$g(x,y) = \begin{cases} 1 & \text{if } f(x,y) >= T \\ \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Every point $(x, y)$ for which $f(x, y) >= T$ is then called an *object point*, whereas all the other ones are said to be *background points*.

### 2.1.1 Challenges in Image Thresholding

The most significant issue with thresholding is the fact that only the intensity is considered and no relationship between pixels. This can lead to the inclusion of external pixels that are not part of the original region. Similarly, isolated pixels can be lost from a region. The presence of noise in the image can easily worsen the outcome, because pixels intensity may not represent the normal intensity in the region. The usage of thresholding is often based on experimentation and small adjustments. Still, a large portion of a region may be lost or the area may be extended with extraneous background pixels (especially when shadows of objects are presents, causing them to be included as part of a dark object on a light background).

One of the disadvantages of global image thresholding is also the fact that it is not particularly effective when changes in illumination occur in the image. This drawback can partially be mitigated by determining thresholds locally. Therefore, instead of a single global threshold value, the threshold itself can actually vary across different parts of the image.

## 2.2 Image Filtering

Filtering is a technique for modifying or enhancing an image. It is used either to emphasize some features of the image or to remove some other. Images can be filtered with various Low-Pass Filters (LPF) or High-Pass Filters (HPF). HPFs are useful for detecting image edges whereas LPFs are typically applied image blurring for removing noise.

### 2.2.1 Correlation and Convolution

These are basic operations that can be applied in order to extract information from an image. In a sense, they are the simplest operations that can be performed but, nevertheless, extremely powerful and useful. Due to their simplicity, they are also well understood, easy implementable, and efficiently computable. Both of these operations fulfill two features, namely:

- linearity, and
- shift in-variance.

Given a square filter with an odd number of elements represented by a $(2N+1)$ x $(2N+1)$ matrix $F$ ($N \in \mathbb{N}$) and the image matrix $I$, the results of correlation can be computed by aligning the center of the filter with a pixel. Overlapping values are then multiplied together

and summed up to make the result corresponding to that given pixel. It can be written as:

$$(F \otimes I)(x, y) = \sum_{i=-N}^{N} \sum_{j=-N}^{N} F(i, j)I(x + i, y + j) \tag{2.2}$$

Convolution is very similar to correlation, but the filter is flipped both horizontally and vertically beforehand:

$$(F \star I)(x, y) = \sum_{i=-N}^{N} \sum_{j=-N}^{N} F(i, j)I(x - i, y - j) \tag{2.3}$$

### 2.2.2 Image Blurring

To achieve a smoothing effect, the image is convolved with a LPF kernel. This particular type of filtering removes high-frequency content from the image (for example noise). Unfortunately, edges also can be blurred a bit while applying this operation, although there are also smoothing filters that prevent edges from being blurred. The OpenCV library provides a user with various filters with the most popular ones are:

- Averaging, it takes the average of all the pixels under kernel area;
- Gaussian, instead of box filter, a Gaussian kernel is used;
- Median, in this case, the median value of all the pixels under kernel area is used; and
- Bilateral, slower compared to other filters, but it keeps edges sharp.

### 2.2.3 Finding Contours

A contour is a list of points that represent, in one way or another, a curve in an image. They come in handy for analyzing the shape of an object, for its detection and for recognition. For better results, it is good to use binary images, so the process of contours finding should be applied after thresholding or edge detecting.

### 2.2.4 Edge Detection

An edge is a place of a sudden discontinuity in an image, which can arise from surface normal, surface color, depth, illumination, or other discontinuities. This rapid change in the image intensity function can be observed in places where the first derivative of this function has local maxima. The primary steps taken in the edge detection process include:

- Smoothing derivatives to suppress noise and compute gradient,
- Threshold to find regions of "significant" gradient,
- Thinning to get localized edge pixels, and
- Connecting edge pixels.

### 2.2.5 Discrete Wavelet Transform (DWT)

DWT basically acts as low-pass and high-pass filter. When an image of size $(M \times N)$ passes through DWT, it yields four different coefficients. Approximate coefficients of level 1 $A_1$ are obtained by applying a low-pass filter to both the horizontal row and vertical columns of pixels. Detailed horizontal coefficients of level 1 $(Dh_1)$ are obtained by applying low-pass filter to the horizontal rows of pixels and high-pass filter to the vertical columns of pixels. Detailed vertical coefficients of level 1 $(Dv_1)$ are obtained due to high-pass filter to the horizontal rows of pixels and low-pass filter to the vertical columns of pixels. Detailed diagonal coefficients of level 1 $(Dd_1)$, due to high-pass filter to both horizontal rows and vertical columns of pixels, are also extracted in the process. Similarly, $Dh_2, Dv_2, Dd_2$, and $A_2$ are the resultant matrices of the second level of 2D-DWT.

A wavelet basis bi-orthogonal wavelet was selected to decompose the fundus images in [15] based on the energy values in the various sub-bands are used as features as shown below:

$$Energy_{sub-band} = \frac{1}{(M \times N)} \sum_{x=M} \sum_{y=N} (D_{x,y}^{sub-band})^2 \tag{2.4}$$

In any image $(x_i, y_i)$ where $i = 1, 2, \ldots, K$ represents the edge points of an object, while the Fourier descriptors of its edge can be represented by the following approach [16]. Each point can be treated as a complex number such that:

$$s(k) = x_i + jy_i \tag{2.5}$$

and the discrete Fourier transform (DFT) of $s(k)$ is:

$$a(u) = \sum_{k=0}^{k-1} s(k) e^{-j2\pi \frac{uk}{K}}. \tag{2.6}$$

If the length of DFT of any sequence is the same as an original sequence, the number of the descriptors varies as the length of the edge changes. Here, the power coefficient of Fourier descriptors is computed as follows

$$PAC = \sum_{u \neq 0, v \neq 0} (F_R^2(u, v) + F_1^2(u, v)) \tag{2.7}$$

where, $F_R(u, v)$ and $F_R(u, v)$ are real and imaginary segments of the Fourier transforms of the image, and $u$ and $v$ are the frequencies along the $x$ and $y$ axes of the image respectively. Fourier descriptors are not invariants to scaling and translation.

### 2.2.6 Histogram-Based Method

These methods are used to represent image regions by encoding texture information of its local descriptors. One such method is the Binarized Statistical Image Features (BSIF) [17] that takes an image and a set of filters and bit size values as input parameters to produce

a textural description of the image. For a given bit size $b$ and filter length $f$, the BSIF method extracts textural representations from a set of images $img$ of size $n$ such that;

$$BSIF(f, b, n, img) = D_{p,n} \tag{2.8}$$

where $p = 2^f, p$ are textural feature representations (dimensions) for each image and $n$ are the number of instances in the data-set $D$. The new number of columns returned after a BSIF textural extraction is proportional to the filter length $f$ parameter passed during the extraction process. For example, processing a images using the BSIF method with filter length 5 (f=5) returns a representation of the image with 23 textural features $2_5 = 32$. We use the BSIF method to extract information from a set of diagnostic images and present the information in a group of multi-dimensional datasets with varying dimension sizes by using a set of bit sizes $b_{set}$ and filter lengths $f_{set}$. The resulting clusters of datasets of varying dimensional sizes per cluster.

## 2.3 Feature Extraction

The textural representation of an images can be obtained from Textural Filter Methods that perform different measurements on the gray-scale intensity of a single or subset of image elements in a localized region of the image. These filter methods are classified as either performing structure extraction operations like distance measure to describe macro structures of the image, or statistical operations like gray-scale correlation and frequency domain analysis to describe its statistical properties of an image.

Traditional texture descriptors are Local Binary Pattern (LBP), Laws Texture Energy (LTE), Entropies, Hus invariant moments [18]. In their study, they employed two popular techniques namely, Discrete Wavelet Transform (DWT) [19], and the Fourier

spectrum [16]. They then extracted the entropy coefficients from these descriptors and used these features to classify images.

Hybrid models like the Gabor Filter [20]–[22] and Haralick Features Method [23]–[25] has recently gained popularity in diagnostic classification of images in the medical field because of the advantages of performing both structures and statistical operations.

### 2.3.1 Binarized Statistical Image Features (BSIF)

Binarized statistical image features (BSIF) is a local image descriptor that produces binary codes for each pixel in the thyroid ultrasound image [15]. The pixels in the image are described as binary-coded strings, and every individual pixel is assigned a code value that represents the intensity distribution in the image. We then use the image histogram to identify texture properties within individual sub-region of the image. Given a neighborhood of $n \times n$ pixels with a set of $x$ linear filters of the same size, the $x$-bit label value of every bit in the binary code string is obtained by binarizing with a linear filter with zero thresholds. This is shown in Equation (2.9).

$$s = Wl \tag{2.9}$$

In the Equation (2.9), $l$ is the $n^2 \times n$ vector representation of the $n \times n$ neighborhood, and $W$ is the $x \times n^2$ matrix denoting the stack of the vector notations of the filters. Every single bit is related with a different filter, and the required length of the bit string sets the number of filters. Likewise, Equation (2.10) shows that the $i^{th}$ value of s is the function of the $i^{th}$ linear filter $w_i$.

$$s_i = w_i^T l_i \tag{2.10}$$

The different filters are learned from a training set of image patches by maximizing the statistical independence of the filter responses. Hence, the BSIF descriptors learn and recognize the filters by independent component analysis (ICA) filters. Each bit of the BSIF value can be attained through Equation (2.11).

$$B_i = \begin{cases} 1 & \text{if } s_i > 0 \\ \\ 0 & \text{if } s_i <= 0 \end{cases} \tag{2.11}$$

The two parameters in the BSIF descriptor (filter size and the length of the binary code string) can be varied. In this work, the BSIF images are passed through 7 filter sizes $\{5, 7, \cdots, 17\}$ and 8-bit length $\{5, 6, \cdots, 12\}$.

### 2.3.2 Haralick Textural Features

Haralick textural features [23] are features calculated based on spatial dependencies of a Gray Level Co-occurrence Matrix (GLCM) in the region of interest (ROI). Texture features obtained from the Haralick method have been used to classify abnormalities in lung CT images as either having tumors or a buildup of fluid. Tables 2.1 to 2.4 explain in detail the notations and equations applied to obtain Haralick Textural Features.

## 2.4 Image Classifiers

Classification is a predictive technique that determines the correct target class of a discrete dependent variable. The most common type of classification technique is the binary classification where there are two possible outcomes. Multi-label classification technique is characterized with having multiple outcomes for a single target variable, while Multi-Target classification technique has two or more target variables with binary or multiple values.

**Table 2.1** Haralick texture features notations.[23]

| **Notations** | **Definition** |
|---|---|
| Element $i, j$ | $x(i,j)$ |
| Number of Gray levels | $N$ |
| $p(i,j)$ | $\frac{x(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} x(i,j)}$ |
| $p_x(i)$ | $\sum_{j=1}^{N} p(i,j)$ |
| $p_y(j)$ | $\sum_{i=1}^{N} p(i,j)$ |
| $\mu_x$ | $\sum_{i=1}^{N} p_x(i)$ |
| $\mu_y$ | $\sum_{j=1}^{N} p_y(j)$ |
| $\sigma_x^2$ | $\sum_{i=1}^{N} (i-\mu_x)^2 \cdot p_x(i)$ |
| $\sigma_y^2$ | $\sum_{j=1}^{N} (j-\mu_y)^2 \cdot p_y(j)$ |
| $p_{x+y}(k)$ | $\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)$ |
| $p_{x-y}(k)$ | $\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)$ |
| $\mu_{x+y}$ | $\sum_{k=2}^{2N} k \cdot p_{x+y}(k)$ |
| $\mu_{x-y}$ | $\sum_{k=0}^{N-1} k \cdot p_{x-y}(k)$ |
| HX | $-\sum_{i=1}^{N} p_x(i) log(p_x(i))$, where $k = i + j$ |
| HY | $-\sum_{i=1}^{N} p_y(j) log(p_y(j))$, where $k = \lvert i - j \rvert$ |
| HXY | $-\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j) \cdot log(p(i,j))$ |

**Table 2.2** Haralick texture features notations (continued).[23]

| **Notations** | **Definition** |
|---|---|
| HXY1 | $-\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)log(p_x(i)\cdot p_y(j))$ |
| HXY2 | $-\sum_{i=1}^{N}\sum_{j=1}^{N} p_x(i)\cdot p_y(j)\cdot log(p_x(i)\cdot p_y(j))$ |
| $Q(i,j)$ | $\sum_{k=1}^{N}\frac{p(i,j)p(j,k)}{p_x(i)p_y(k)}$ |

The task of a classification algorithm is to learn a model function that maps a non-empty set of observations to one or more outcomes/targets. However, the number of observations/features observed in medical studies are usually larger than the number of samples size. This could result in performance degradation of the classification algorithms due to the curse of dimensionality phenomenon. Robust supervised machine learning algorithms have been developed to process datasets with a significantly small sample size relative to the large number of features. Methods like the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) and their variants have been successfully applied to high dimensional data; however, these require the probabilities of the observed output to be known a priori.

Traditional classification techniques like Support Vector Machine, Random Forest and Linear Regression and Logistic Regression Analysis techniques are used when the class probability is not known a priori, which is the case for most real world domains. Classification and regression analysis techniques are supervised learning techniques that are used to predict the value of an output variable given a set of real or categorical input values. Their choice of application is determined in terms of target output to

**Table 2.3** Gray-level co-occurrence matrix (GLCM) notation[23].

| **Features** | **Definition** |
|---|---|
| Entropy | $-\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)log(p(i,j))$ |
| Energy | $\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)^2$ |
| Difference Entropy | $-\sum_{k=0}^{N-1} p_{x-y}(k)log(p_{x-y}k)$ |
| Sum Entropy | $-\sum_{k=2}^{2N} p_{x+y}(k)log(p_{x+y}(k))$ |
| Homogeneity | $\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{p(i,j)}{1+(i+j)^2}$ |
| Inverse Difference | $\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{p(i,j)}{1+\lvert i-j\rvert}$ |
| Max. Probability | $\max_{i,j} p(i,j)$ |
| Sum Average, $\mu_{x+y}$ | $\sum_{k=2}^{2N} kp_{x+y}(k)$ |
| Sum of Squares | $\sum_{i=1}^{N}\sum_{j=1}^{N} (i-\mu)^2 p(i,j)$ |
| Cluster Prominence | $\sum_{i=1}^{N}\sum_{j=1}^{N} (i+j-2\mu)^3 p(i,j)$ |
| Contrast | $\sum_{i=1}^{N}\sum_{j=1}^{N} (i-j)^2 p(i,j)$ |
| Cluster Shade | $\sum_{i=1}^{N}\sum_{j=1}^{N} (i+j-2\mu)^4 p(i,j)$ |
| Dissimilarity | $\sum_{i=1}^{N}\sum_{j=1}^{N} \lvert i-j\rvert \cdot p(i,j)$ |
| Sum Variance | $\sum_{k=2}^{2N} (k-\mu_{x+y})^2 p_{x+y}(k)$ |
| Correlation (Cor) | $\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{i-\mu_x}{\sigma_x}\frac{j-\mu_y}{\sigma_y} p(i,j)$ |
| Max Cor. Coefficient | $\sqrt{\lambda_2 Q(i,j)}$ |

**Table 2.4** Gray-level co-occurrence matrix (GLCM) notation (continued)[23].

| Features | Definition |
|---|---|
| Info. Measure of Cor 1 | $\frac{HXY - HXY1}{max(HX, HY)}$ |
| Info. Measure of Cor 2 | $\sqrt{1 - exp[-2(HXY2 - HXY)]}$ |
| Autocorrelation | $\sum_{i=1}^{N}\sum_{j=1}^{N}(i \cdot j)p(i,j)$ |
| Difference Variance | $\sum_{k=0}^{N-1}(k - \mu_{x-y})^2 p_{x-y}(k)$ |

be estimated. While classification techniques are used to predict discrete target output, Regression Analysis fit a model that estimates a target output to a continuous set of value. An independent sample set is introduced to the model to evaluate performance.

As more observations are A Regression problem can be transformed into a classification problem during the pre-processing steps by discretizing the continuous dependent variables into a set of intervals but it is not recommended to transform a classification problem into a Regression Analysis problem due to the infinite and unordered nature of continuous dependent variables.

### 2.4.1 Logistic Based Classifiers

Du, X., et al. [26] evaluated eight classification methods for the classification of epileptic electroencephalogram (EEG) signals from epilepsy. In their study, the Regression Model achieved the highest True Positive and Area Under Curve rates for the classification of the “before seizure” class when a relatively high Principal Components Analysis (PCA) percentage was applied.

### 2.4.2 Rule-Based Classifiers

Rule-based classifiers have gained considerable interest in decision support systems because of their ease of interpretability [27]. Dua, S., et al., [28] proposed a rule-based data-adaptive shrinking approach with Partial Decision Trees (PART) to classify high dimensional physiochemical properties of protein structures. In their study, they hypothesize that feature descriptors can be deprived through measurement of their sparseness in the feature domain. Singh, H., et al., in [29] proposed a quantization-based dimensionality reduction technique for mining Association rules in their work. They identify isomorphic relationships of protein physiochemical properties to segment and classify a multi-domain protein family.

Jang, et al., [30] suggested the use of an Adaptive-Network-Based Fuzzy Inference System (ANFIS). It uses the fuzzy modeling procedure whereby it learns information about a data-set from the data-set itself. Here the membership functions are determined from the given set of features and rules are generated by adjusting the weights using both the forward pass and back-propagation algorithms.

Fuzzy-based computer-aided diagnosis (CAD) techniques were used in the diagnosis of diabetic retinopathy [31] for feature extraction and classification while Nayak et al. [32] proposed the application of textural features in conjunction with a feed-forward Artificial Neural Network (ANN) for the classification of diabetic Maculopathy. Nayak et al. [32] reported in their findings a sensitivity of 95.4% and a specificity of 100%. Similarly, the work presented by Chowriappa et al. [33] proposed an ensemble selection technique to build an ensemble classifier from a library of models. They reported an average accuracy of 96.7%, with the highest recorded accuracy of 97.8%.

#### 2.4.2.1 Bayesian Classifiers

Rule-based Bayesian Network models provide the means to capture distributions structures of real-life phenomena, that are difficult to model without expert knowledge of the domain. They can also be used to learn BN model structures from raw data if the domain structure is unknown. The process of learning a BN model is divided into two phases, the parameter learning and structure learning phases.

Sahu, S., et al., [34] proposed a Bayesian approach for modeling a speckle removal algorithm for noise reduction and edge preservation in medical ultrasound images. They suggested using a Cauchy prior to model the true wavelet while the Gaussian Probability Density Function was used to model the noisy coefficients. An optimized Bayesian Least Square Estimation (BLSE) approach in the work presented by Nagaraj, Y., et al., [35] was proposed for detecting and denoising of speckles in ultrasound images of the carotid artery. This is beneficial for estimating thickness of tunica intima and tunica media and predicting risk of the cardiovascular disease. Related works [36], [37] in fundus classification rely on novel features extraction techniques to classify images to their corresponding classes. These features describe morphological characteristics of the fundus image such as cup-to-disc (c/d) ratios [36], diameter of the optic disc, the distance between the optic disc center and Optic Nerve Head (ONH), and ratio of area of blood vessels in inferior-superior side to the area of blood vessels in the nasal-temporal side[32].

In recent years, textural features have gained prominence in fundus image classification. Enhancement to the extraction of textural features for classification is described in the work by Acharya et al., [37]. They proposed the use of texture and higher-order

spectra features for the classification of glaucoma using fundus images. They classified fundus images into normal and glaucoma classes and achieved a classification accuracy of 91%. Similarly, the work by Dua et al. [36] also proposed the use of wavelet-based energy features on fundus images to classify glaucomatous images. They reported an overall accuracy of 93.33%.

# CHAPTER 3

# RESEARCH METHODOLOGY

In this study, we propose an adaptive convolution kernel filtering method that suppresses speckle noise and enhances edges of objects detected in the image using a deterministic approach of obtaining the kernel filter. We address the challenge of learning kernel filter parameters as a supervised learning task and address the challenge of extracting relevant features in a high-dimensional and sparse dataset. The main task in regression analysis is to understand the relationship between the independent and continuous dependent variables by fitting a predictive model that minimizes the error of the predicted and actual output.

Our method uses a trained Regression model to approximate optimal values for the kernel values. The Haralick method is then applied to extract features that represent texture characteristic of the image. We propose an image filtering method the adapts to information obtained from the image domain using a trained regression model. The proposed method summarized in Figure 3.1 consists of the following steps:

1. Image Pre-processing
2. Filter Design
    (a) Regression Model Design
    (b) Estimate Kernel Parameters

3. Feature Extraction

4. Model Evaluation

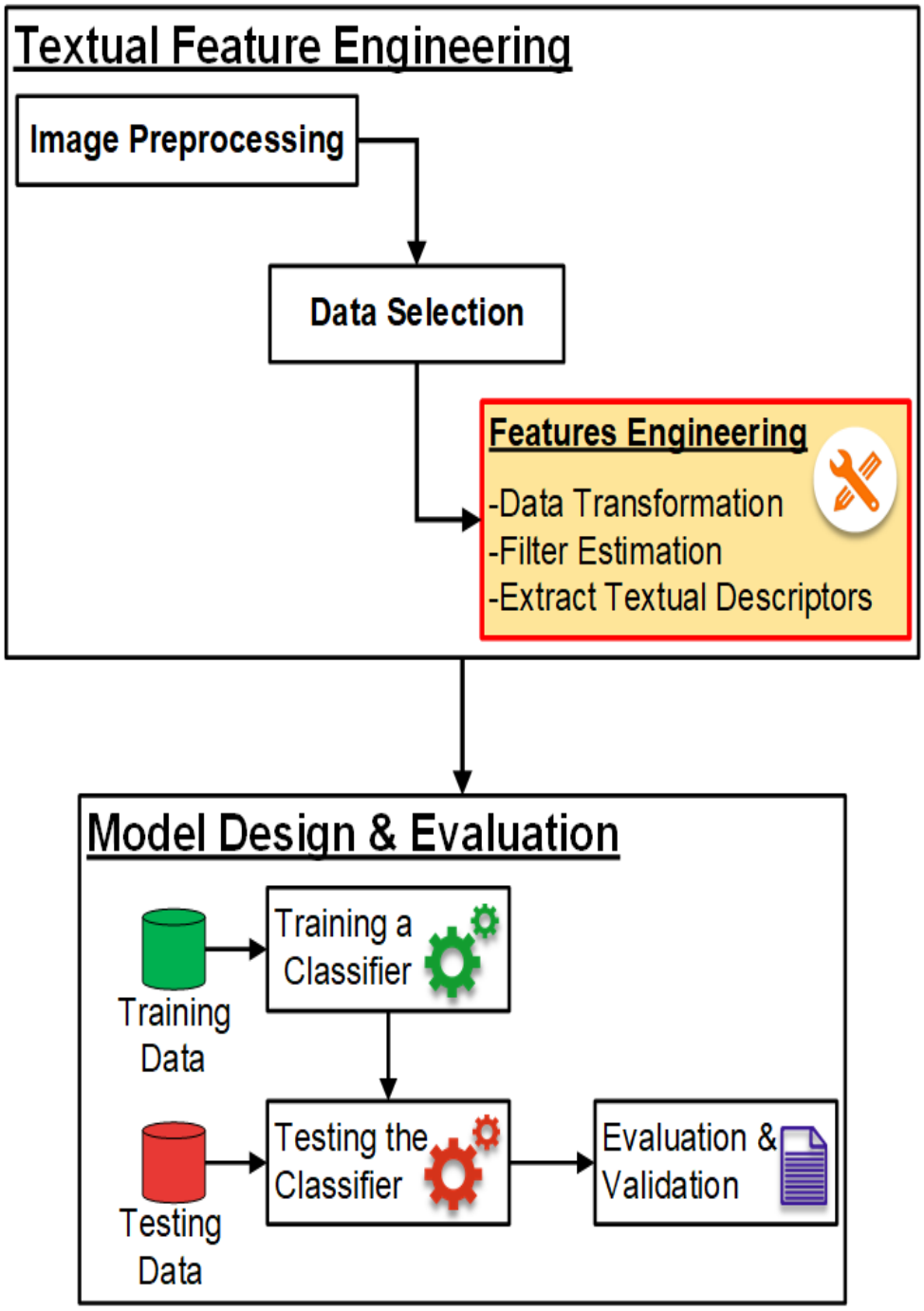



**Figure 3.1** The Block Diagram for the proposed Computer-Aided Diagnosis (CAD) System for the automated distinction of benign and malignant thyroid lesions

## 3.1 Research Approach

Ultrasonography as a non-invasive diagnostic technique has gained popularity as a safe diagnostic process due to the low operational costs and non-radiation exposure involved [38]. The drawback to this process, however, is that air and bone tissues tend to distort the sound waves rendering the image to be very noisy. A study of linear and non-linear filtering methods like the Discrete Wavelet Transform (DWT) [39] methods for suppressing speckle noise in ultrasound images demonstrated a significant improvement in the quality of the images. An evaluation of thyroid and carotid artery images showed a significant relationship between higher kurtosis and filtered images when filtered with 16 different de-speckling methods [40].

## 3.2 Data Preparation

The dataset after the extraction of BSIF consisted of fifty-six individual BSIF files. Each file represented a filter size and bit length combination, consisting of 344 samples. The number of attributes for each file varied. Diagrammatic presentation of the dataset after performing the BSIF extraction method is illustrated in Figure 3.2 and Figure 3.3.

**Definition 3.1.** *A data set $D$ represented by a set of matrices $x_{(i,j)}^{(b\times f)}$ with each individual matrix of size $n$ is defined as;*

$$D = \{X, Y\} \tag{3.1}$$

*where* $X = \left\{ x_{(i,j)}^{(b,f)} \mid i \right.$ *represents a sample* $(1 \leq i \leq n)$; $j$ *represents the BSIF attributes* $\forall\ (b, f)$; *with bit length* $b \in B,\ B = \{5, \cdots, 12\}$, *filter size* $f \in F,\ F = \{5, 7, 9, 11, 13, 15, 17\} \Big\}$, *and* $y_i \in \{0, 1\}$

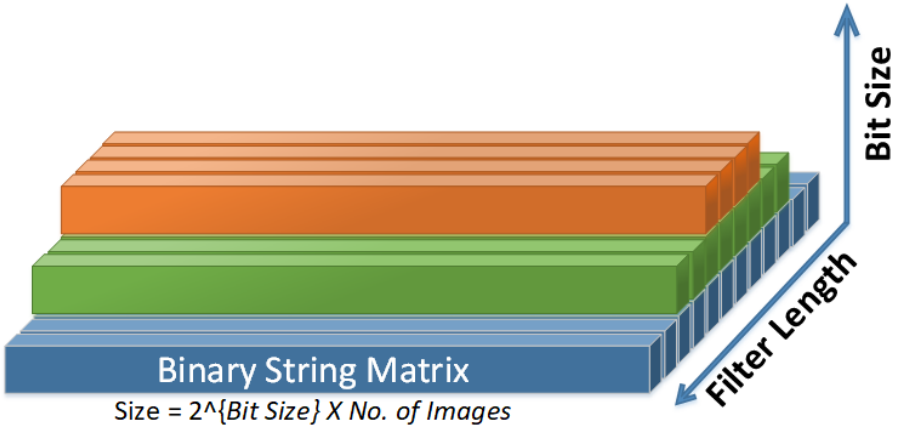



**Figure 3.2** A Multilevel dimensional dataset using 3 BSIF filter levels

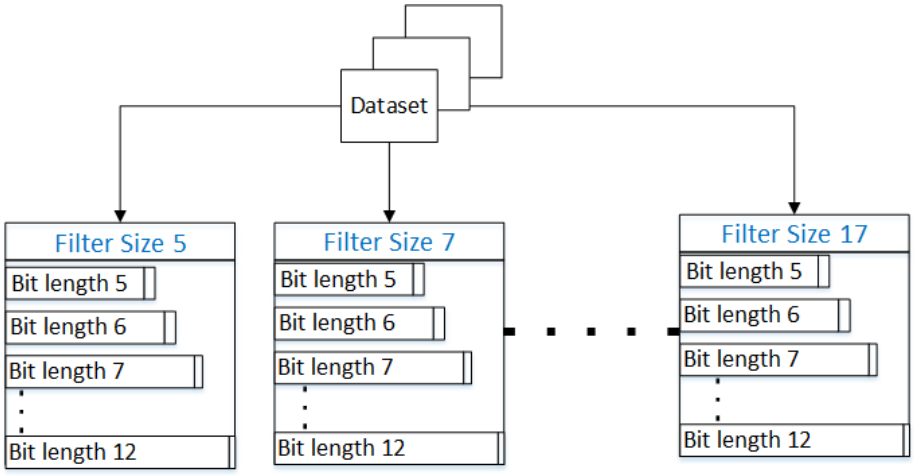



**Figure 3.3** Data matrix in the BSIF features space

### 3.3 Image Pre-processing

During pre-processing, attenuation artifacts within the ultrasound images were handled applying a Contrast Limited Adaptive Histogram Equalization (CLAHE) [41], [42]. We believe that the CLAHE method will enhance the contrast, thereby improving the resolution quality and reduces noise representation in the image [43]. An inverse transformation is performed to enhance detection of geometric information in the spatial domain.

## 3.4 Proposed Image Filter Design

The selected sampled image patch was transformed into a 1-dimensional structured dataset with the target response represented by the expected pixel value after applying (b) high-pass filtering within regions of interest (ROI) and (c) low-pass filtering on out-of-scope regions (non-ROI) region. A linear regression model is fitted to the transformed data set to estimate model coefficient parameters. An Elastic-Net regularization path was used to control the concentration and shrinkage of coefficient parameters during the model training process.

### 3.4.1 Regression Model

A Regression Model was used to model the relationship between independent input variables and their dependent target variables by fitting an equation. The association between the input and target variables can be determined by first examining the data points onto a scatter-plot and analyzing their correlation coefficient. A positive correlation coefficient value indicates that there exists a positive association between the variables and a negative values for the correlation coefficient indicates a negative association. The regression coefficient values can be estimated using singular values decomposition.

#### 3.4.1.1 Linear Regression

A Linear Regression assumes the residual error are random follows a normal distribution for input variables $X$ and dependent variables $Y$ such that the model equation is defined as:

$$\begin{aligned} Y_i &= \beta_0 + \beta X_{i,*} + \epsilon_i \forall j \in 1, \ldots, p \\ &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \end{aligned} \tag{3.2}$$

where $i$ are the number of observations for $i = 1, \ldots n$, $j$ are the features for $j = 1, \ldots p$, $\beta = (\beta_1, \ldots, \beta_p)$ are coefficient parameter $\beta$ and $\epsilon_i \sim N(0, \sigma^2)$ are the residual error parameters with a probability density function:

$$f(y_i) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} exp\left[\frac{-\left(y_i - X_{i,*}\right)^2}{2\sigma^2}\right] \tag{3.3}$$

The prediction of $Y_i$ of the regression model is equal to:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y \tag{3.4}$$

#### 3.4.1.2 Singular Value Decomposition (SVD)

Singular Value Decomposition proposed in [44] is a matrix decomposition method commonly used in dimensionality reduction in high dimensional and strongly collinearity matrix. The singular value decomposition of a matrix $A$ is defined as:

$$A = UDV^T \tag{3.5}$$

where $D$ an $n \times n$ diagonal matrix with the singular values, $U$ an $n \times n$ matrix of left singular vectors, $V$ a $p \times n$ matrix of right singular vectors, and columns of $U$ and $V$ are orthonormal, i.e:

$$UU^T = I = VV^T \tag{3.6}$$

The Equation (3.4) can be rewritten in terms of the SVD-matrices in Equation (3.5) to estimate coefficient parameters $\hat{\beta}$ of the linear regression model as:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X + \lambda I)^{-1} X^T Y \end{aligned} \tag{3.7}$$

where $\lambda$ is the tuning parameter (Ridge penalty parameter), $\lambda \in [0, \infty)$ and $I$ is a $(p \times p)$-dimensional identity matrix. The Newton-Raphson equation is applied iteratively to find the roots of the log-likelihood estimation function in Equation (3.2) to derive new values of the coefficient as:

$$\begin{aligned}\hat{\beta}_{new} &= \hat{\beta}_{old} + (X^T W X)^{-1} X^T \left[Y = g^{-1}(X;\beta)\right] \\ &= (X^T W X)^{-1} X^T W \left\{ X\hat{\beta}_{old} + W^{-1}\left[Y - g^{-1}(X;\beta)\right] \right\} \\ &= (X^T W X)^{-1} X^T W Z \\ &= \arg\max_{\beta} (Z - X\beta)^T X^T W (Z - X\beta)\end{aligned} \tag{3.8}$$

where $W$ is a $(n \times n)$-dimensional diagonal matrix with $W_{i,i} \in [0, 1]$ representing the weight of the $i^{th}$ observation, $Z$ is the adjusted response $Z = X\hat{\beta}^{old} + W^{-1}[Y - g^{-1}(X_{i,*}; \beta^{old})]$. The Link function $g(;)$ is a function that links responses to observed variables $X_{i,*}$ such that:

$$p(X_{i,*}) = p(Y_i) = g(X_{i,*};\beta)^{-1} = \frac{exp(X_{i,*};\beta)}{1 + exp(X_{i,*};\beta)} \tag{3.9}$$

The likelihood estimates of the Linear Regression is thus defined as:

$$L(Y|X;\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp\left[\frac{-(y_i - X_{i,*}\beta)^2}{2\sigma^2}\right] \tag{3.10}$$

and is rescaled to a log-likelihood estimation as:

$$\mathcal{L}(Y|X;\beta) = -n\log\sqrt{2\pi\sigma} - \frac{1}{2\pi\sigma^2}\sum_{i=1}^{n}(y_i - X_{i,*}\beta)^2 \tag{3.11}$$

#### 3.4.1.3 Logistic Regression

Logistic Regression or Logit Regression is a type of probabilistic statistical classification model used to explain the response of binary targets. The Logistic Regression

algorithm iterative applies a Maximum Log-likelihood estimator to obtain a set combination of the coefficients of the observations. The coefficients are the fit to a regression loss function that approximates the probability of the discrete target value. Because Linear Regression functions are unbound, a logarithmic transformation is performed to obtain a Logistic Regression function of the form:

$$Y_i = \beta_0 + X_{i,*}\beta_i \rightarrow \ln\left[\frac{p(X_{i,*}))}{1 - p(X_{i,*})}\right] \tag{3.12}$$

$$= \beta_0 + X_{i,*}\beta_i, \forall p(X_{i,*}) \in [0, 1]$$

where $p(X_{i,*})$ is a logistic link function that links the response values to input variables.

A numerical root-finding method, like the Newton-Raphson method, is used interactively to approximate the coefficient parameters $\beta$ by converging to a maximum log-likelihood estimate. The Likelihood estimation function of the Logistic Regression model is defined as:

$$L(Y|X;\beta) = -\prod_{i=1}^{n}[(p(y = 1|X_{i,*})^{Y_i}(p(y = 0|X_{i,*})^{1-Y_i})] \tag{3.13}$$

The Equation (3.13) is rescaled using logarithms to reduce the likelihood of estimations being skewed towards large values. This is known as the log-likelihood estimation function as $\mathcal{L}(Y|X;\beta)$ denoted as:

$$\mathcal{L}(Y|X;\beta) = -\frac{1}{n}\sum_{i=1}^{n}[Y_i \cdot (\beta_0 + X_{i,*}^T\beta) \tag{3.14}$$

$$+ \ln(1 + exp^{(\beta_0 + X_{i,*}^T\beta)})] \tag{3.15}$$

### 3.4.2 Regularization

The model coefficients $\hat{\beta}$ of high-dimensional data are known to be highly sparse and tend to have an infinite number of possible estimations for a single coefficient. Regularization controls the coefficients by modeling an objective path that reduces both the variance and the Sum of Square Errors (SSE) between the observed and predicted targets. This is achieved by applying a penalty factor to the log-likelihood function of a Logistic Regression model as denoted below:

$$\mathcal{L}_*(Y|X;\beta) = \max_{\beta_0,\beta}\{\mathcal{L}(Y|X;\beta) + Obj(P;\beta)\} \tag{3.16}$$

where $Obj(P;\beta)$ is the objective function for setting a penalty parameters $P$.

#### 3.4.2.1 Ridge Regularization

The Regularization paths proposed to handle such challenges include the Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net Regularization Paths. The Ridge Regularization Path shrinks correlated coefficients towards a common estimated value and its objective function is denoted as:

$$\mathcal{L}_{Ridge}(Y|X;\beta) = \mathcal{L}(Y|X;\beta) + \lambda \cdot \frac{1}{2}\sum_{j=1}^{p}\beta_j^2 \tag{3.17}$$

#### 3.4.2.2 Least Absolute Shrinkage and Selection Operator Regularization(LASSO)

The Least Absolute Shrinkage and Selection Operator (LASSO) Regularization path eliminates irrelevant coefficients from the model its objective function is defined as:

$$\mathcal{L}_{LASSO}(Y|X;\beta) = \mathcal{L}(Y|X;\beta) + \lambda \cdot \sum_{j=1}^{p}|\beta_j| \tag{3.18}$$

where $\lambda \in \mathbb{R}^+$ set to control the $\sum_{j=1}^{p}|\beta_j|$ (LASSO) $L1$ normalization and $\sum_{j=1}^{p}\beta_j^2$ (Ridge) $L2$ normalization.

#### 3.4.2.3 Elastic Net Regularization

The Elastic Net Regularization sets a balance between the Ridge and LASSO paths by introducing of a trade-off parameter $\alpha, 0 \leq \alpha \leq 1$. The path taken by the Elastic Net will be similar to that of a Ridge Path if $\alpha$ is equal to 1 and a path similar to LASSO if $\alpha$ is equal to 0. The log-likelihood function for the Elastic Net Regularization is denoted as:

$$\mathcal{L}_{ENet}(Y|X;\beta) = \mathcal{L}(Y|X;\beta) + \lambda \cdot [(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha \cdot \sum_{j=1}^{p}|\beta_j|] \tag{3.19}$$

#### 3.4.2.4 Coefficient Shrink Factor

In addition to the Ridge and LASSO factors, a rate of shrink factor of each coefficient can be set to the objective function $Obj(P_{\lambda,\alpha};\beta)$ to control the penalty ratio of each coefficient individually. The Objective function in an Elastic Net Regularization model can be transformed into a weighted Objective function as:

$$Obj(P_{\lambda,\alpha};\beta) = \lambda\sum_{j=1}^{p}\omega_j[(1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|] \tag{3.20}$$

where where $\omega_j$ is the rate of shrink of coefficient $j$, for $0 \leq \omega_j \leq 1$

### 3.4.3 Estimating Kernel Parameters

Variations of the filter kernel are generated from the estimated model coefficients that were achieved from adjusting the $L_1$ and $L_2$ penalties by applying Elastic Net regularization and minimizing the root square error.

A set of Kernels were obtained from sampling different image patches sizes $(i.e., 5, 7, \cdots, 29)$ corresponding to the sizes of the resultant kernels. The resulting filters were applied to the images to extract Haralick texture features illustrated in Figure 3.4.

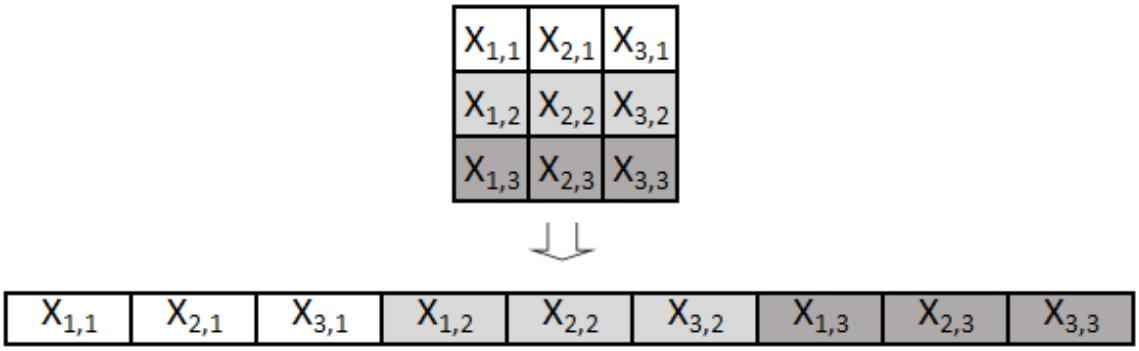



**Figure 3.4** Transformation of image patches to a 1-D data frame

### 3.4.4 Spectral Analysis

In this work, the ultrasound images are transformed from an image domain to a spectrum domain in order to analyze their natural repeating patterns. The Spectral analysis techniques this we use in this work is a Fourier Series technique. Fourier Series is an approximation of a complex periodic function as simplified infinite sums of cosine and sine functions for the purpose of the analyzing the function in another domain other than its original domain. In order to model a function as Fourier Series, we assume that the sampled periodicity of the function is bound by a finite non-zero interval and repeats in either time or space or both time and space like in wave motion.

A 1-dimensional function $f(x)$ is said to be periodic if it repeats over a period $x \in [\frac{-L}{2}, \frac{L}{2}]$ within the interval $L$ if $f(x+L) = f(x), \forall x \in \mathbb{R}$. The function $f(x)$ can be modeled as a Fourier Series of the form;

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\left[a_n cos\left(\frac{nx\pi}{L}\right) + b_n sin\left(\frac{nx\pi}{L}\right)\right] \tag{3.21}$$

with the where $a_0$, $a_n$ and $b_n$ are Fourier coefficients denoted as;

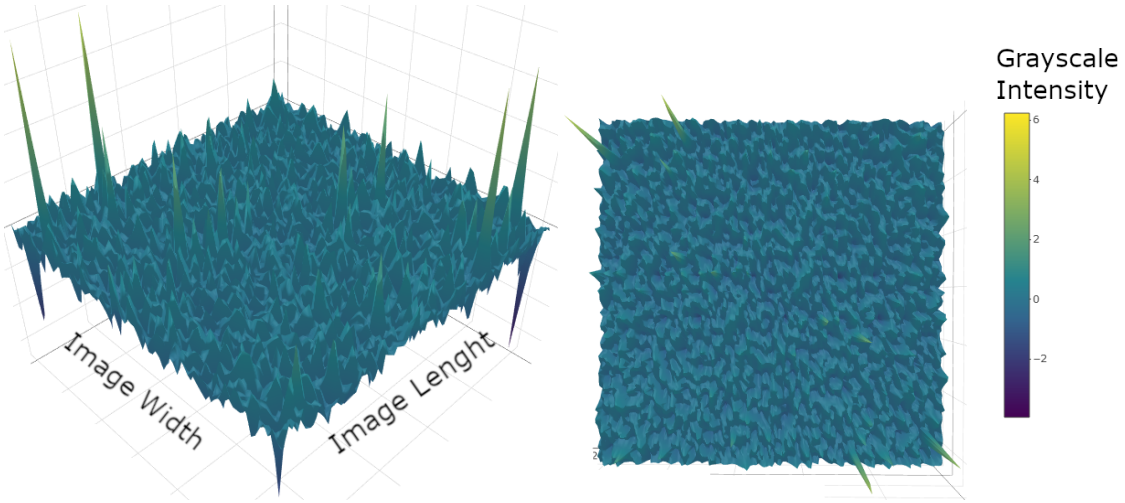
$$a_0 = \frac{1}{\pi}\int_{\frac{-pi}{2}}^{\frac{pi}{2}} f(x)dx \tag{3.22}$$

$$a_n = \frac{1}{\pi} \int_{\frac{-pi}{2}}^{\frac{pi}{2}} f(x) cos\left(\frac{nx\pi}{L}\right) dx \tag{3.23}$$

$$a_n = \frac{1}{\pi} \int_{\frac{-pi}{2}}^{\frac{pi}{2}} f(x) sin\left(\frac{nx\pi}{L}\right) dx \tag{3.24}$$

In order to represent Equation (3.20) as a Fourier Series of the form of Equation (3.21), we assume that for a fixed set of Ridge and LASSO regularization penalty terms, there exists a continues weighted coefficient terms $\omega_j$ with a periodic interval of $[0, 1]$ and that $\omega_j$ represents the degree of magnitude for the coefficients $\beta_j$. We also assume that the log-likelihood function is bound by a maximum estimate of $Obj(P_{\lambda,\alpha}; \beta)$. We propose an analysis of the Fourier Series of the Ultrasound Images (Figure 3.5) as an expansion of the Logistic Regression problem as;

$$Obj(P_{\lambda,\alpha}; \beta) = f(\omega) = \lambda \sum_{j=1}^{p} \omega_j \left[ (1 - \alpha) \frac{1}{2} \beta_j^2 + \alpha \left| \beta_j \right| \right] \tag{3.25}$$




**(a)** Surface view of Fourier coefficients **(b)** Heat map plot of Fourier coefficients

**Figure 3.5** 2-D gray-scale intensity of ultrasound image after a fourier transformation.

## 3.5 Feature Engineering

Our next task was to select features that are not highly correlated. We adopted a two-phase approach, namely, ***Feature Extraction*** to extract domain-relevant features and later ***Feature Selection*** on the extracted features and to select features that maintain and/or contribute information towards the domain classification problem. In the Feature Selection phase, the Principal Component Analysis (PCA) was applied alongside the Boruta Feature Selection Algorithm to perform dimensionality reduction and provide scalability of the proposed model.

### 3.5.1 Feature Extraction

In the first phase, we applied the Haralick Texture Feature Extraction method and evaluated the classification performance of the extracted features against traditional image feature extraction methods like the Binarized Statistical Image Features (BSIF).

#### 3.5.1.1 Haralick Texture Features

The Haralick texture features method [23] was used in this work to identify and extract texture features of the ultrasound image. Haralick texture features are statistical calculations of an image or region of interest (ROI) in the image based on spatial dependencies of a Gray Level Co-occurrence Matrix (GLCM) using four angular mean. This approach assumes that spatial dependence frequencies between neighboring cells in a gray scale image are a function of the angular relationship and their distance as illustrated in Appendix F. Geometrical relationships of GLCM measurements are applied to four radian angles $\theta = 0, \pi/4, \pi/2, \pi/4$ and the distance $d = max\{|x|, |y|\}$ to adjacent neighboring cells.

### 3.5.2 Feature Selection

The goal of feature selection in classification is to select of a subset of features that when excluded from the dataset would not affect the accuracy performance of the classifier. The features are known to have minimal predictive information relative to other features in the same sample set. Feature Selection methods are categorized in three major techniques, namely, ***Filter Selection***, ***Wrappers*** and ***Embedded*** techniques.

***Filter Selection Techniques*** select a subset of features based on statistics methods like the Chi-Square Test [45] and Information Gain [46] to determine their relevance while ***Wrapper*** methods determine the subset to be selected by measuring their importance to a classifier obtained from a learning machine. Examples of wrappers include Recursive Feature Elimination [47] and Sequential Forward Selection [48], [49]. ***Embedded*** techniques are similar to ***Wrappers*** in that they too use learning methods to evaluate feature relevance; however, this process is incorporated in the training model design. An example of a algorithm that applies the ***Embedded*** technique for feature selection is the Boruta method.

#### 3.5.2.1 Feature Selection with Principal Component Analysis (PCA)

The principal components are computed by the Eigen-decomposition technique that assigns each component a weighted linear composite of the original attributes. The resultant weighted feature vectors represent the ordered variance retained by the components corresponding to the original data.

**Definition 3.2.** *After PCA, the transformed data set $\hat{D}$ is defined as,*

$$\hat{D} = \left\{ \lambda_{(i,j)}, Y_i \right\} \tag{3.26}$$

*where* $\lambda_{(i,j)} = \left\{ \lambda_{(i,j)}^{(b \times f)} \mid \lambda_{(i,j)} \right.$ *represents the principle components of the attribute matirx of* $\hat{x}_{(i,j)} \; \forall \; (b \times f) \Big\}$

### 3.5.2.2 Boruta Feature Selection Algorithm

The Boruta algorithm [50], is a wrapper based feature selection algorithm that uses the Random Forest (RF) ensemble classifier. In this method, illustrated in Figure 3.6, features are selected in random subsets and each subset is evaluated based on a derived measure of importance referred to as the Z Score. Given a random forest with $M$ trees

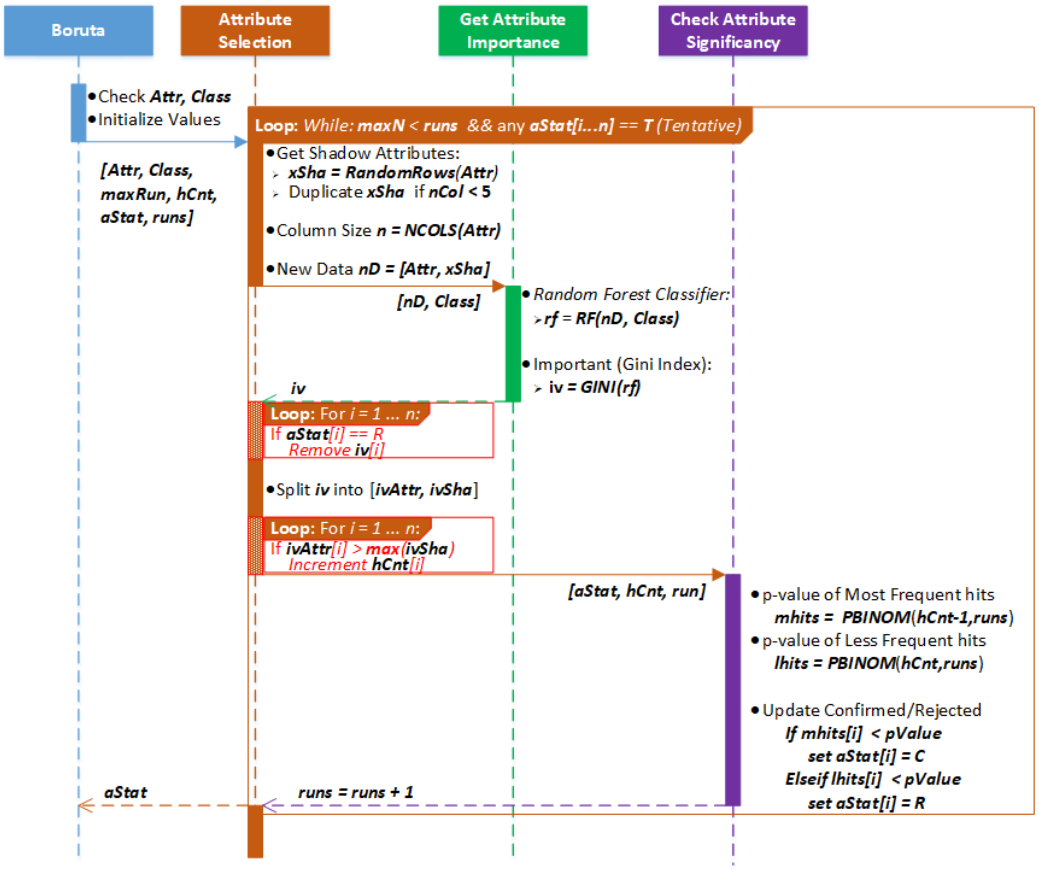



**Figure 3.6** System architecture diagram of the Boruta feature selection algorithm

and a set of features $\lambda_{i,j}$ in each tree $T_m$ for $1 \leq m \leq M$, the Z-score $(ZS)$ for feature is

defined as:

$$ZS = \frac{da_m - MDA_m}{\sigma_m} \tag{3.27}$$

where $da_m$ and $\sigma_m$ and $MDA_m$ are the decrease of accuracy, standard deviation and Mean Decrease Accuracy (MDA) of each feature $\lambda_{i,j}$ in tree $T_m$.

The MDA is computed as:

$$MDA_m = \frac{1}{M} \sum_{k=1}^{M} da_k \tag{3.28}$$

# CHAPTER 4

# BAYESIAN BASED CLASSIFICATION

The Bayes' Theorem (also known as Bayes' Rule) was formulated by the British statistician and Royal Society Fellow Thomas Bayes (1701 - 1761). He proposed a theorem for solving problems where there are two or more competing hypotheses, given prior knowledge of the data. After his death, his work was carried on by Richard Price (1764) [51] and popularized by Pierre-Simon Laplace [52] in 1774.

## 4.1 Bayesian Networks (BN)

A Bayesian Networks (BN) is a probabilistic graphical model representing a set of variables and their conditional dependence through a directed acyclic graph (no directed cycles) and a set of probability distributions for each variable. A BN for a set of random variables is built by computing the Joint Probability distribution of all their possible outcomes and then applying the simple Markov Chain rule to extract conditional probabilities [53]. The Markov Chain rule states that if the current probability of observing an event or state of a node is known, there is no need to compute its previous probability in order to determine what the future probability will be. This simple rule is applied to the construction of a Directed Acyclic Graph (DAG) structure assuming a node is independent to all other nodes on condition that they are not descendants of its parent node.

In this section, we introduce terms that are frequently used when explaining a probabilistic models.

A Marginal Probability (MP) $p(\boldsymbol{x}_A)$ is the unconditional probability of an event $\boldsymbol{x}_A$ to occur, thus there is no conditional assumption or dependency from another event for the current event $\boldsymbol{x}_A$ to occur.

The formula for the MP is defined as;

$$\boldsymbol{p}(\boldsymbol{x}_A) = \int p(\boldsymbol{x}_A, \boldsymbol{x}_B) d\boldsymbol{x}_B \tag{4.1}$$

where the integration is carried out only on the domain of random variables $\boldsymbol{x}_B$, $\boldsymbol{x}_B \subseteq \mathbb{R}$.

The Conditional Probability of $\boldsymbol{x}_A$ is the Conditional Probability of an event $\boldsymbol{x}_A$ given that another event $\boldsymbol{x}_B$ occurs. The formula for Conditional Probability of $p(\boldsymbol{x}_A|\boldsymbol{x}_B)$ is defined as;

$$p(\boldsymbol{x}_A|\boldsymbol{x}_B) = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{p(\boldsymbol{x}_B)} \tag{4.2}$$

given that $p(\boldsymbol{x}_B) > 0$ and $p(\boldsymbol{x}_B) \neq 0$.

The Joint Probability $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ is the probability of both events A and event B occurring. The Joint Probability distribution of events in Figure 4.1 can be represented as:

$$\begin{aligned} p(\boldsymbol{x}_A, \boldsymbol{x}_B, \boldsymbol{x}_C, \boldsymbol{x}_D) &= p(\boldsymbol{x}_A) \times p(\boldsymbol{x}_B|\boldsymbol{x}_A) \\ &\times p(\boldsymbol{x}_C|\boldsymbol{x}_A) \times p(\boldsymbol{x}_D|\boldsymbol{x}_B, \boldsymbol{x}_C) \\ &\times p(\boldsymbol{x}_E|\boldsymbol{x}_D) \end{aligned} \tag{4.3}$$

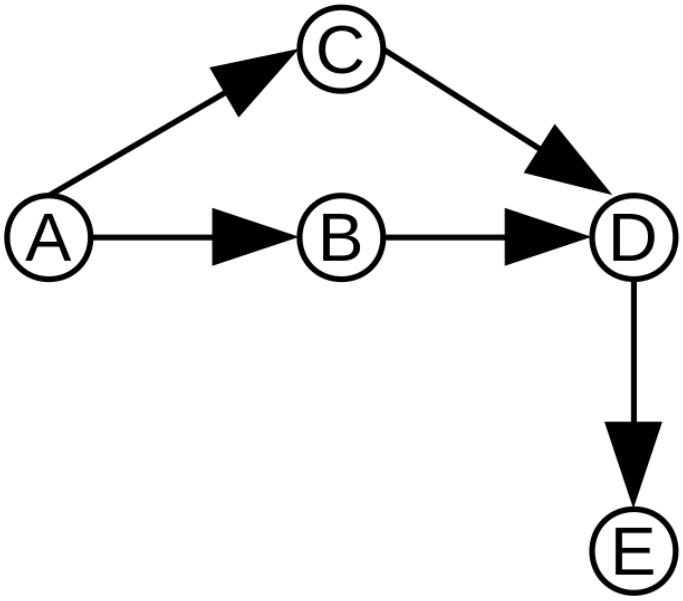



**Figure 4.1** Illustration of a joint probability distribution for events: A, B, C, D and E

## 4.2 Markov Chains

A ***Markov Chain*** is a state-wise algorithm used to infer the uncertainty of parameters of a complex non-linear mechanical and probabilistic problems through a Markov Network. In Bayesian Networks, a Hidden Markov Chain can be used to solve the scalability challenges of naive inference algorithms. Unlike the Markov Chain where the transition information for each state in the Markov Network is visible for all, state information in a Hidden Markov Chain are not directly visible to the observer. However, the probability distribution of each state to transition to another state is available to the observer.

A Hidden Markov Chain on a state space $\mathcal{S}$ is defined as a sequence of random variables $\{\boldsymbol{x}_{(i)}; i \geq 0\}$ where the states represent successive iterations, such that the Conditional Probability distribution of $\boldsymbol{x}_{(i+1)}$ follows the Markov assumption that:

- the current state is conditionally independent of all the other past states given the state on the previous time steps; and
- the evidence at a time $t_s$ depends only on the state $\mathcal{S}$.

The Markov Chain is defined by its Joint Probability;

$$p\left(\boldsymbol{x}_{(i+1)} \,|\, \boldsymbol{x}_{(i)}, \boldsymbol{x}_{(i-1)}, \ldots, \boldsymbol{x}_{(0)}\right) = p\left(\boldsymbol{x}_{(i+1)} \,|\, \boldsymbol{x}_{(i)}\right) \tag{4.4}$$

where $\mathcal{S}$ is the state space of all possible outcomes of the random variables $\boldsymbol{x}_{(i)}$ and $p\left(\boldsymbol{x}_{(0)}\right)$ is the initial MP distribution of $\boldsymbol{x}_{(0)}$ at state $i = 0$

As an example of a discrete-state Markov chain (see Figure 4.2), consider a 3-state Markov chain representing changes of weather condition with $S = \{\text{Rainy}, \text{Sunny}, \text{Cloudy}\}$ and a transition probability matrix $P$ such that:

$$\begin{aligned} P &= \begin{pmatrix} p(\text{Rain} \text{ — } \text{Rain}) & p(\text{Sun} \text{ — } \text{Rain}) & p(\text{Cloudy} \text{ — } \text{Rain}) \\ p(\text{Rain} \text{ — } \text{Sun}) & p(\text{Sun} \text{ — } \text{Sun}) & p(\text{Cloudy} \text{ — } \text{Sun}) \\ p(\text{Rain} \text{ — } \text{Cloudy}) & p(\text{Sun} \text{ — } \text{Cloudy}) & p(\text{Cloudy} \text{ — } \text{Cloudy}) \end{pmatrix} \\ &= \begin{pmatrix} 0.15 & 0.20 & 0.65 \\ 0.22 & 0.63 & 0.15 \\ 0.20 & 0.05 & 0.75 \end{pmatrix} \end{aligned} \tag{4.5}$$

The Bayesian Network Structure in Figure 4.3 is used to represent the uncertain domains of the nodes $A$, $B$, $C$, $D$ and $E$ from their Joint Probability distribution

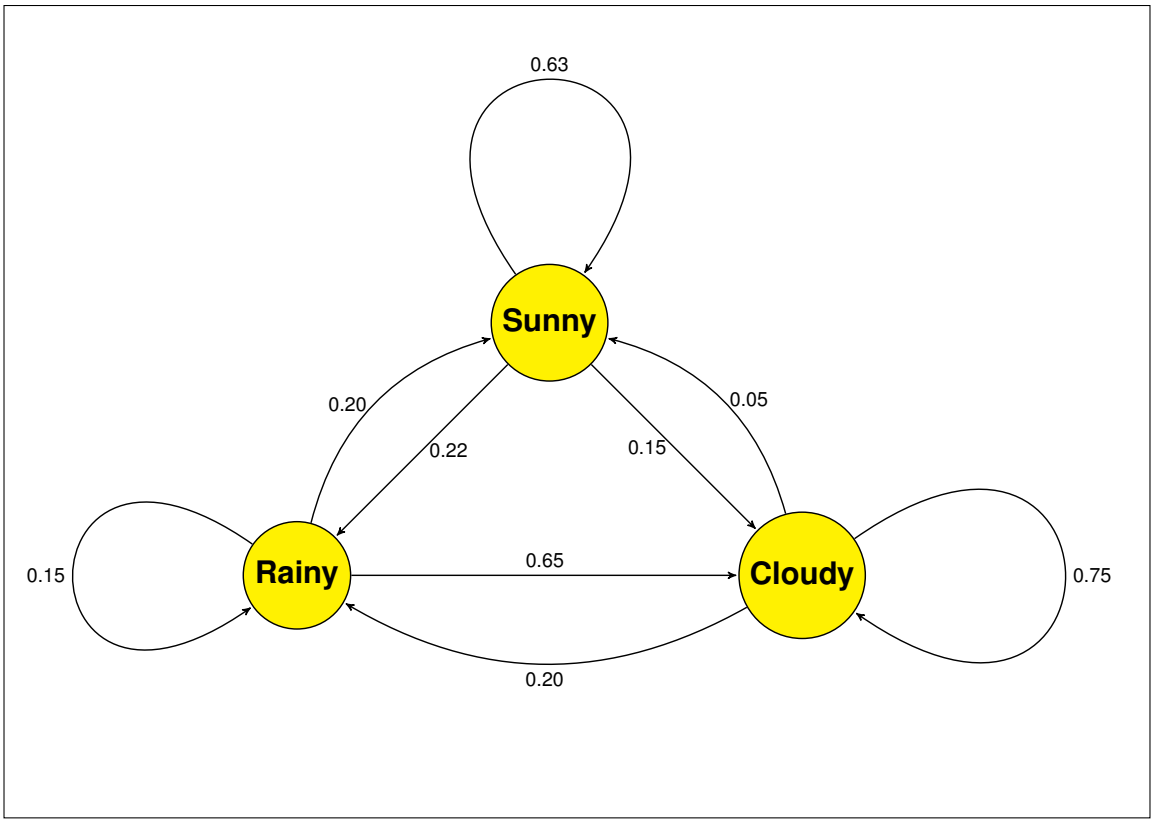
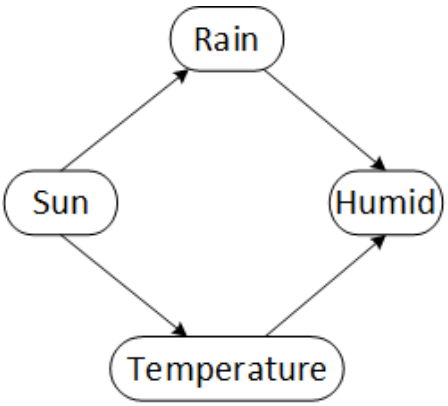



**Figure 4.2** Graph representation of a 3-state Markov Chain

$P(A, B, C, D, E)$. The directed arrows are drawn between nodes to describe their conditionally (or unconditionally) independent relationships.

## 4.3 Bayesian Classifiers

Bayesian Classifiers are a kind of Machine Learning Classifiers that apply Bayes' Rule to build a Probabilistic Model that predicts the class membership of a target output given a set or subset of the features based on assumptions between the features. The Probabilistic Model applies Bayes' Rule to estimate posterior probabilities of the feature from deriving prior probabilities and likelihood estimates.




| Sun | Rain=Yes | Rain=No |
|---|---|---|
| Yes | 0.40 | 0.60 |
| No | 0.65 | 0.55 |

| Sun=Yes | Sun=No |
|---|---|
| 0.28 | 0.72 |

| Sun | Temp=High | Temp=Low |
|---|---|---|
| Yes | 0.8 | 0.20 |
| No | 0.5 | 0.50 |

| Rain | Temp | Humid=Yes | Humid=No |
|---|---|---|---|
| Yes | High | 0.67 | 0.33 |
| Yes | Low | 0.34 | 0.66 |
| No | High | 0.63 | 0.37 |
| No | Low | 0.10 | 0.90 |

**Figure 4.3** A Bayesian inference network illustrating the conditional probability of the weather forecast

A key challenge in learning a Bayesian classifier from data is selecting an optimal network structure that minimizes the posterior expected loss. This is equivalent to maximizing the information gain in Decision Tree Algorithms in Equation (4.8).

For a given data set $D$ with $n$ independent and identically distributed ($i.i.d.$) events, the goal of Bayesian Theorem is to estimate a ***prior*** probability distributions $\theta$, where $\theta$ is the probability for any of the events to occur prior to observing the data. Inferences about $\theta$ (***posterior*** distribution) can be made by considering its conditional distribution after observing the data. We denote variables or attributes with $x$ while their values are denoted with $x$. The ***posterior*** distribution $p(\theta|x)$ for the observed data $x \in (x_1, \ldots, x_n)$ can be expressed with Bayes' rule as:

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{p(x_{1:n})} \propto p(x_{1:n}|\theta)p(\theta) \tag{4.6}$$

$$p(\theta|x) = p(x_1|\theta) \times p(x_2|\theta) \times \cdots \times p(x_n|\theta) \times p(\theta)$$

where $p(\theta)$ is the prior distribution before any data is observed, $p(x|\theta)$ is the likelihood that is conditional on $\theta$, and $p(x)$ is the evidence or marginal likelihood of the observed data. Its general expression is derived as:

$$P(x) = \prod_{x_i}^{n} P(x_i|Px_i) \tag{4.7}$$

where $Px_i$ are the parents of node $x_i$ in Bayesian Network $BN$. Information Gain (or Entropy Reduction) formula in Equation (4.8) is used to evaluate the influence each attributes $x$ contributes to affect the class probability.

$$\log_2 p(\theta|x) - \log p(\theta) \tag{4.8}$$

#### 4.3.0.3 Parameter Learning

The principle aims of parameter learning in BN models is to estimate probability distribution that explains constraints and uncertainty/subjective belief in the data given available and newly observed evidence. This is a contradiction to the Generative models that use a frequentist approach of estimating the data distribution for a fixed parameter $\theta$ from the observed data.

#### 4.3.0.4 Structure Learning

In the absence of expert knowledge, BN models can be used to learn the graph structure of the network from the data. These learning tasks are primarily categorized into three types with respect to the their algorithm design, namely Constraint-based, Score-based and Hybrid Learners.

Constraint-based Structure Learners apply conditional independence tests to build the BN Structure through an Inductive-Causation Algorithm (Verma and Pearl, 1991). Examples of such learners algorithms include the Grow-Shrink (GS) [54], Incremental Association algorithms (IAMB) [55] and the Fast-Incremental Interleaved Incremental Association algorithms (Fast-IAMB)[56].

Alternatively, Score-Based Structure Learner use heuristic greedy algorithms that determine a fitness score used for implementing the BN. These include the Hill-Climbing (HC) Algorithms [57] and a variation of HC known as Tabu Search (TABU) [58]. TABU Algorithm addresses the limitation of a local optima experienced in the HC algorithm.

Hybrid Structure Learners; however, apply both Constraint-based and Score-based learning techniques to reduce the searchable nodes within a BN structure, while optimizing the network within the reduced structure. Examples of Hybrid algorithms include the Max-Min Hill-Climbing (MMHC) Algorithm [59] and the Hybrid Parents and Children(HPC) Algorithm [60].

## 4.4 Bayesian Model

We developed our Bayesian Model for a multivariate posteriors distribution with Random Effect Factors in Equation (4.9) and adopted the Monte Carlo Markov Chain (MCMC) algorithm proposed by Metropolis and Hastings [61] to estimate the posteriors.

$$y_n = x_n \ + \epsilon_n \tag{4.9}$$

where $\epsilon_n \sim \mathsf{normal}(0, \Sigma)$ and a Probability Mass Function of;

$$\mathrm{F}(y \mid \theta) = \begin{cases} \theta & \text{if } y = 1, \text{ and} \\ 1 - \theta & \text{if } y = 0. \end{cases}$$

We then modify our Model in Equation (4.9) by introducing the random factors from the Feature Engineering process to finally attain a Probability Mass Function as;

$$\mathrm{F}(y \mid x, \alpha, \beta, \gamma) = \prod_{1 \le i \le n} \mathrm{F}(y_i \mid \text{logit}^{-1}(\alpha_i + x_i \cdot \beta))$$

$$= \prod_{1 \le i \le n} \begin{cases} \text{logit}^{-1}(\alpha_i + \sum_{1 \le j \le m}(x_{ij} \cdot \beta_j + \gamma_r)) & \text{if } y_i = 1, \text{ and} \\ 1 - \text{logit}^{-1}(\alpha_i + \sum_{1 \le j \le m}(x_{ij} \cdot \beta_j + \gamma_r)) & \text{if } y_i = 0. \end{cases}$$

where $x \in \mathbb{R}^{n \cdot m}$ are the input variables, $\alpha \in \mathbb{R}^n$ is the intercept, $\beta_j \in \mathbb{R}^m$ are the Fixed effect factors, and $\gamma_r \in \mathbb{R}^k$ for k Random factors. The density calculations of the posteriors estimates distributions returned MCMC Algorithm (see Figure 4.4) are used to further filter our data by eliminating poorly performing factors.

In the study, these factors translate to the filter size, Bit Lengths, and the Alpha Regression parameter $\alpha$ applied in Equation (3.20). We test the wellness of out model using the Hamiltonian Monte Carlo algorithm. As can be seen in Figure 4.5, the divergences in the the marginal energy distribution is indicated with a light blue tail thus establishing whether our model had a challenge during sampling for estimated posteriors.

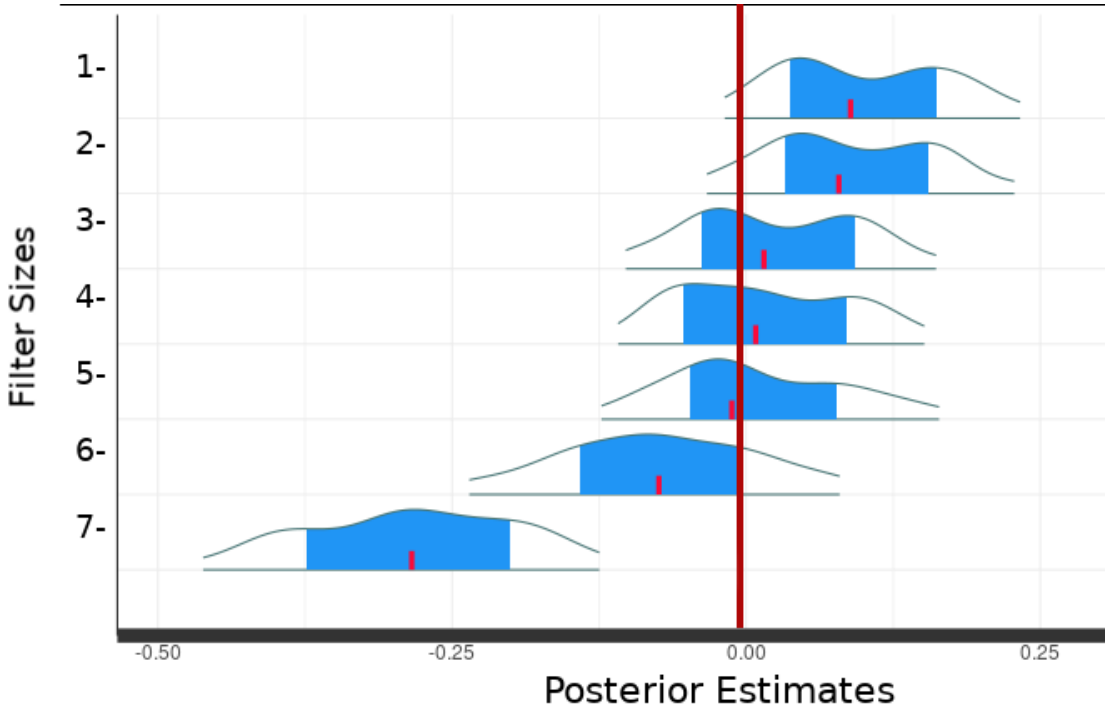



**Figure 4.4** Posterior distributions means of the proposed model using Monte Carlo Markov Chain

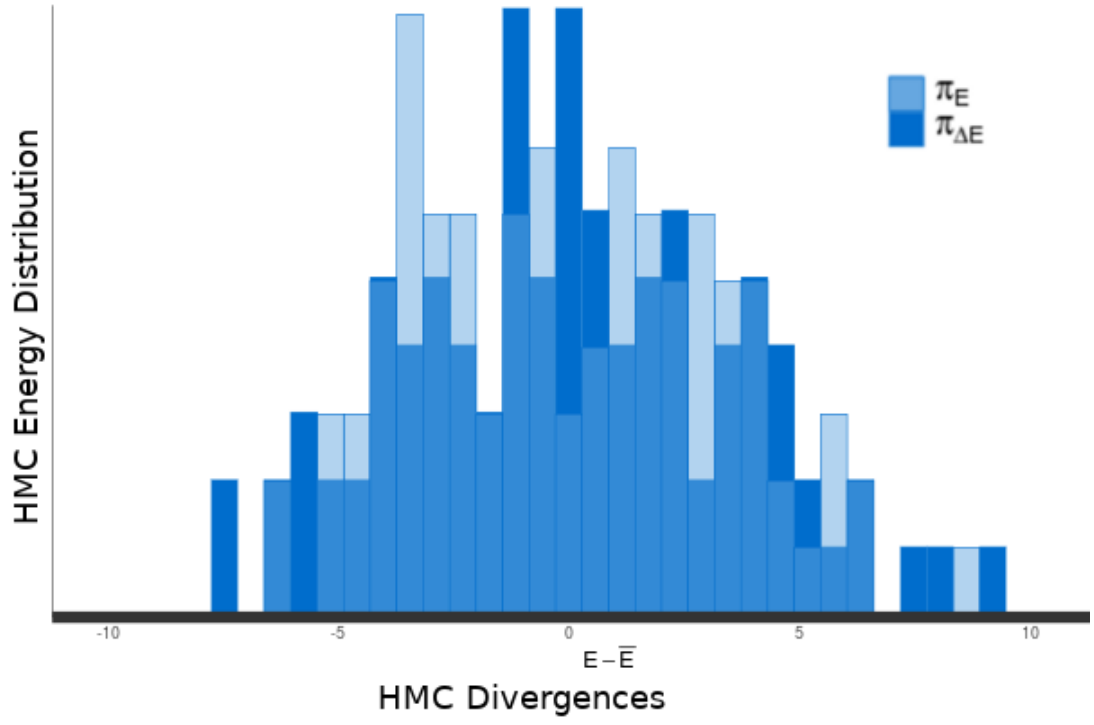



**Figure 4.5** Model performance diagnostic with the Hamiltonian energy divergences

# CHAPTER 5

# EXPERIMENTAL RESULTS

## 5.1 Experimental Data

Ultrasound image data of benign and malignant thyroid lesion used in this work was obtained from the Computer Imaging and Medical Applications Laboratory (CIM@LAB) Figure 5.1. The CIM@LAB is a web-based open access dataset system developed the National University of Colombia - Universidad Nacional de Colombia (UNC) and Instituto de Diagnostico Medico (IDIME) [62]. A total of 388 individual patient cases with 400 diagnostic observation were used in this study. Each observation was stored in two formats, namely:

- an uncompressed JPEG image, and
- an annotated XML file.

Each Patient's case was annotated by expert radiologists and the information is saved in an XML file. This information included age, gender, description of the thyroid nodule, shape, margin, calcification, echogenicity and a Thyroid Imaging Reporting and Data System (TI-RADS) scale described in Table 5.1. The TI-RADS lexicons 1, 2 & 3 were categorized as benign while 4a, 4b, 4c & 5 were categorized as malignant. New cases with respective patient information (as an Ultrasound JPEG image and an XML file) are upload to the database periodically. The images were captured with various ultrasound


**Figure 5.1** A screen shot of the computer imaging and medical applications laboratory system[62].

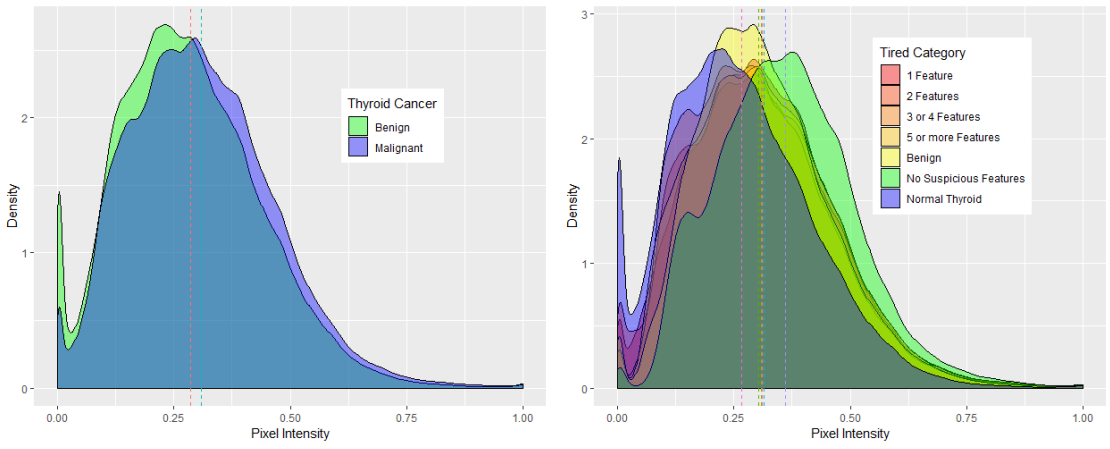
equipment, namely a TOSHIBA Nemio 30, a TOSHIBA Nemio MX, and an Ultrasound linear and convex transducers set at 12MHz. Hence, not at all the ultrasound scans from every subject uses the same ultrasound machine.

We assumed the data followed a Gaussian distribution from the analysis in Figure 5.2 and scaled it to a standard normal distribution with a mean of 0 and a standard deviation of 1. We partitioned the data into a 60/40 split for training set and testing sets. Random samples non-overlapping image patch from a training set were then selected to be fitted

**Table 5.1** The Thyroid Imaging Reporting and Data System (TI-RADS) scale used to describe thyroid lesions

| Scale | Description | Classification |
|---|---|---|
| **1** | Normal thyroid | Benign |
| **2** | Benign Follicular Nodule | Benign |
| **3** | No suspicious lesion | Benign |
| **4a** | 1 suspicious lesion | Malignant |
| **4b** | 2 suspicious lesion | Malignant |
| **4c** | 3 or 4 suspicious lesions | Malignant |
| **5** | 5 or more lesions | Malignant |

to a Linear Regression model. The training set was encapsulated inside a 10-fold cross validation procedure illustrated in Appendix B. Nine folds were fitted to the model and one of the folds was retained for validation of the model. The cross-validation process was then repeated 10 times. Estimated model coefficients at varying $L_1$ and $L_2$ configurations of an Elastic Net regularization were structured into a matrix of filter kernels.




**(a)** A histogram of ultrasound pixel intensity for thyroid cystic nodules

**(b)** Comparison of ultrasound pixel intensity for different Thyroid Imaging Reporting and Data (TIRAD) Categories

**Figure 5.2** Ultrasound images of benign thyroid lesions.

Examples of the XML format for benign and malignant thyroid lesion images are shown in Figure 5.3.a and Figure 5.3.b and examples of the JPEG images for Benign thyroid lesion and Malignant thyroid lesion used in this study are shown in Figure 5.4 and Figure 5.5 respectively.

XML

```xml
<case>
    <number>359</number>
    <age>50</age>
    <sex>F</sex>
    <composition/>
    <echogenicity/>
    <margins/>
    <calcifications>non</calcifications>
    <tirads>2</tirads>
    <reportbacaf/>
    <reporteco>PSEUDONODULES TIROIDITIS </reporteco>
        <mark>
            <image>2</image>
            <svg>[{"points": [{"x": 346, "y": 56}, {"x":
        </mark>
    </case>
```

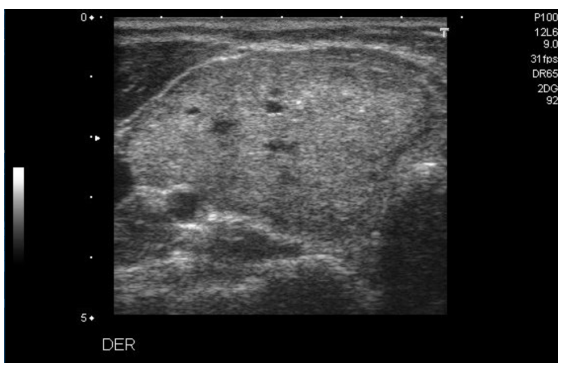
**(a)** Benign follicular lesion

XML

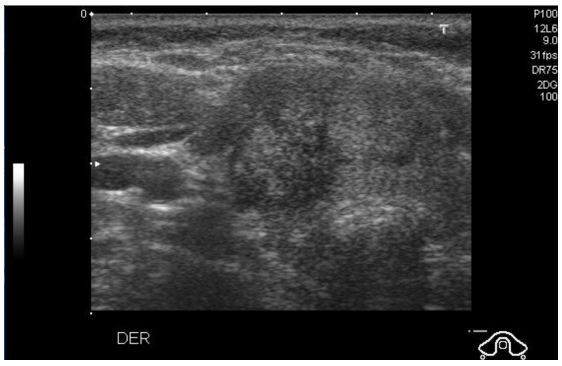
```xml
<case>
    <number>595</number>
    <age>69</age>
    <sex>M</sex>
    <composition>spongiform appareance</composition>
    <echogenicity>hyperechogenecity</echogenicity>
    <margins>well defined smooth</margins>
    <calcifications>microcalcification</calcifications>
    <tirads>4a</tirads>
    <reportbacaf/>
    <reporteco>ARROW : CYSTIC COMPONENT </reporteco>
    <mark>
        <image>1</image>
        <svg>[{"points": [{"x": 303, "y": 126}, {"x": 30
    </mark>
</case>
```

**(b)** 3 or 4 suspicious Malignant lesions

**Figure 5.3** Expert annotations of benign Figure 5.3.a and malignant Figure 5.3.b thyroid lesions [62].

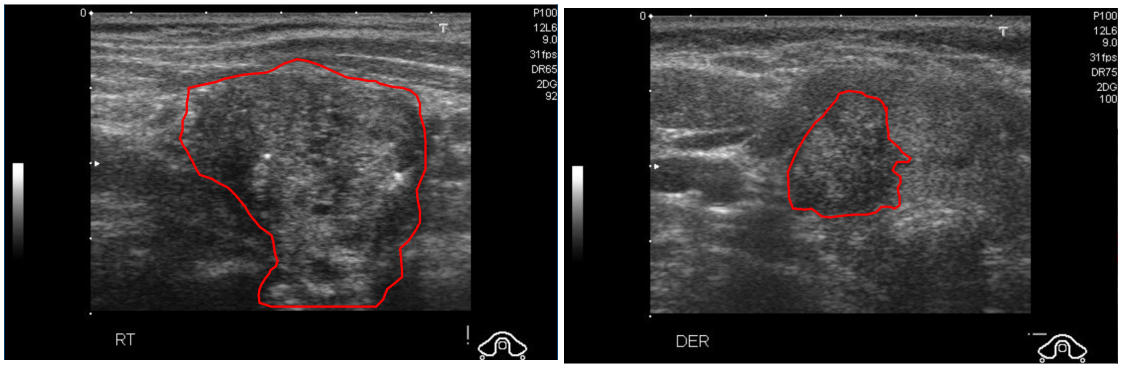



**(a)** Benign follicular lesion




**(b)** Unsatisfactory follicular lesions [62].

**Figure 5.4** Ultrasound images of benign thyroid lesions.




**(a)** 3 or 4 follicular neoplasm **(b)** Malignant

**Figure 5.5** Ultrasound images of malignant thyroid lesions [62].

## 5.2 Image Filter

We select a moving rectangular windows of a specified filter size (see Figure 5.6) and slide it along regions in a 2-dimensional data set. The sampled regions are then transformed into a 1-dimensional structured data set with the region-of-interest as the class label (see Figure 5.7). We then estimate the kernel parameters from coefficients of a Gaussian and Binomial Regression modal using the LASSO and Ridge regularization (see Figure 5.8.a and Figure 5.8.b). We adopt the Elastic Net Regularization model and analyze its weighted objective function when expressed as a Fourier Series (see Figure 5.8.c and Figure 5.8.d).

## 5.3 Feature Selection

### 5.3.1 Kruskal-Wallis Test for Feature Selection

We performed partial dependence test of each selected Haralick features estimated from the single class and multilabel class models Figure 5.9 below. The partial dependence is the approximation of the relationship between a subset of the feature space. Feature importance measures were also computed by applying the Boruta algorithm with the

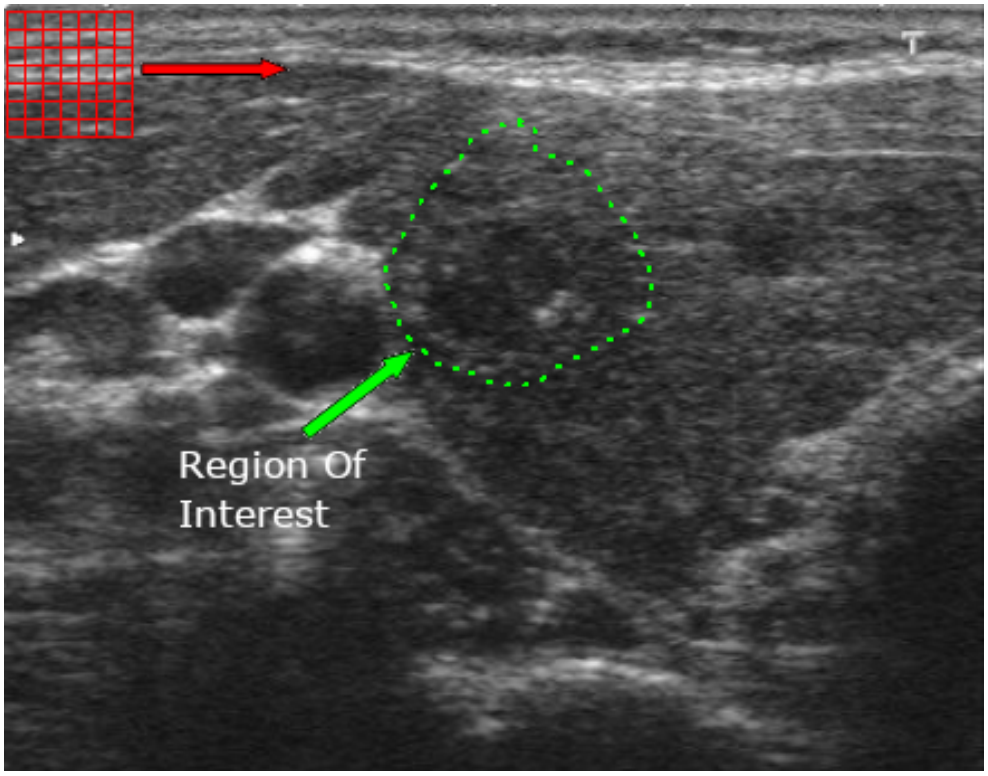



**Figure 5.6** Target region with sliding filter window.

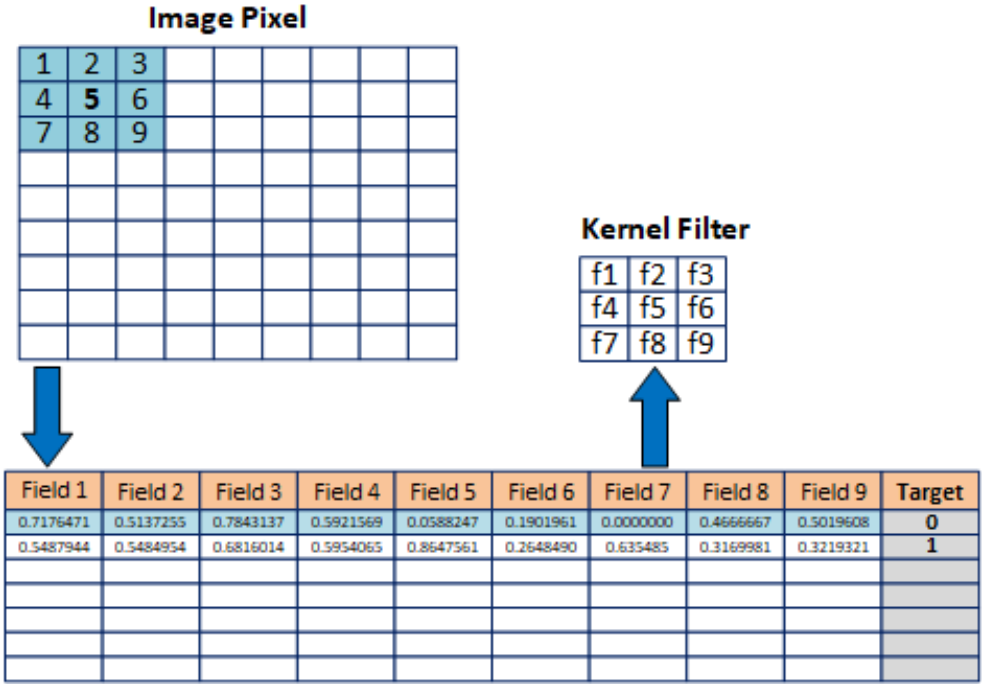
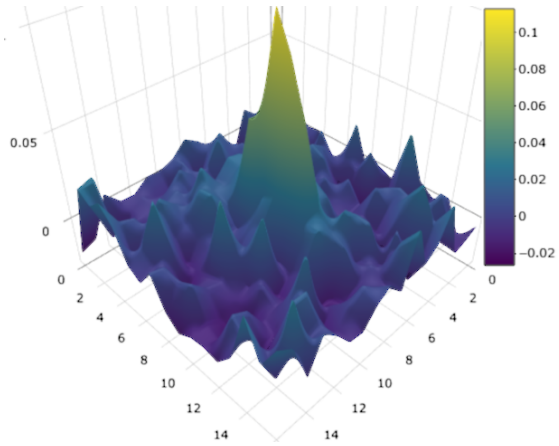
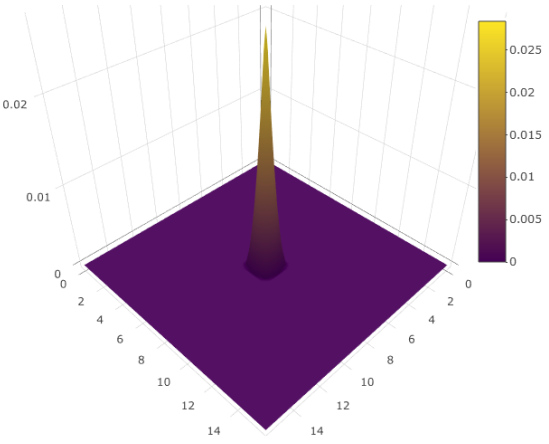



**Figure 5.7** Transformation of image pixel data into a structured data set

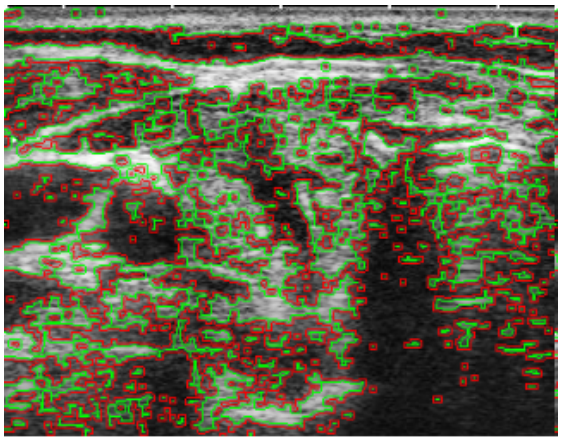
Random Forest (RF) classifier of 500 trees and $\sqrt{n}$ randomly selected features for each tree. On completion, the Boruta algorithm categorized the features into three groups, namely: important, tentative, and unimportant in Figure 5.10. Using a Bayesian network structure, when established conditional dependencies between Haralick features and spatial-independent features (as seen in Figure 5.11), space-based features (as seen in Figure 5.12), and moment-based features (as seen in Figure 5.13). We observed a



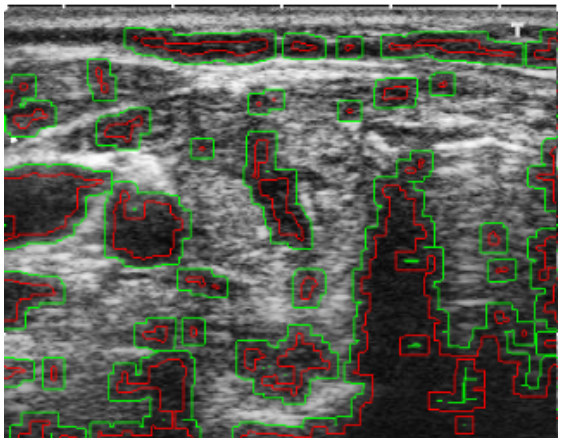

**(a)** LASSO regularization ($\alpha = 1$)




**(b)** Ridge regularization ($\alpha = 0$)



**(c)** Transformed image illustrating detected filter edges



**(d)** Transformed image illustrating detected filter edges

**Figure 5.8** Illustration of kernel filter estimated and their respective filter transformation

higher conditional dependency between Haralick and spatial-independent features than with space-based and moment-based features.

### 5.3.2 Proposed Bayesian Network Structure (BNS)

In our model, we propose to extract Haralick texture features with the feature extraction method introduced in Chapter 3. Components are a good way of explaining the maximum variance using PCA. Prior to applying the PCA algorithm, we used a Box Cox transformation to transform non-normal dependent variables into normal-shape dependent variables for analysis of variance. We fitted our model onto multiple classifiers (illustrated in Table 5.2) and analyzed the performance of our proposed feature

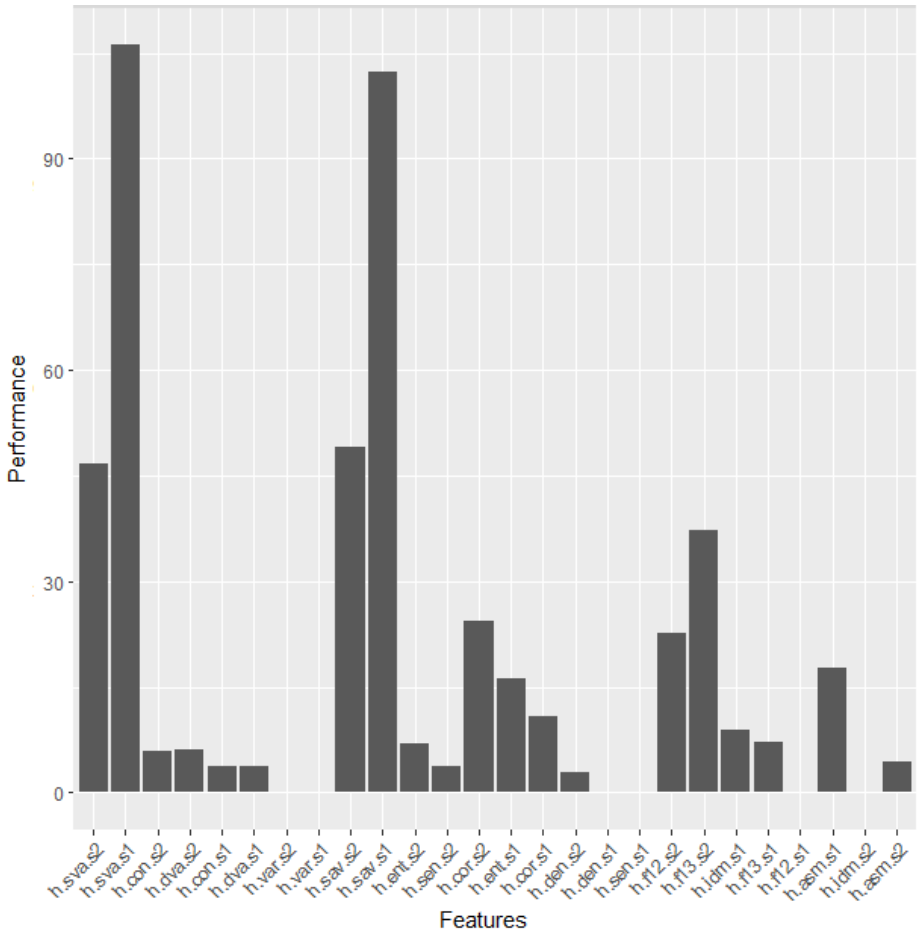



**Figure 5.9** Mean features Kruskal measure of statistical significant

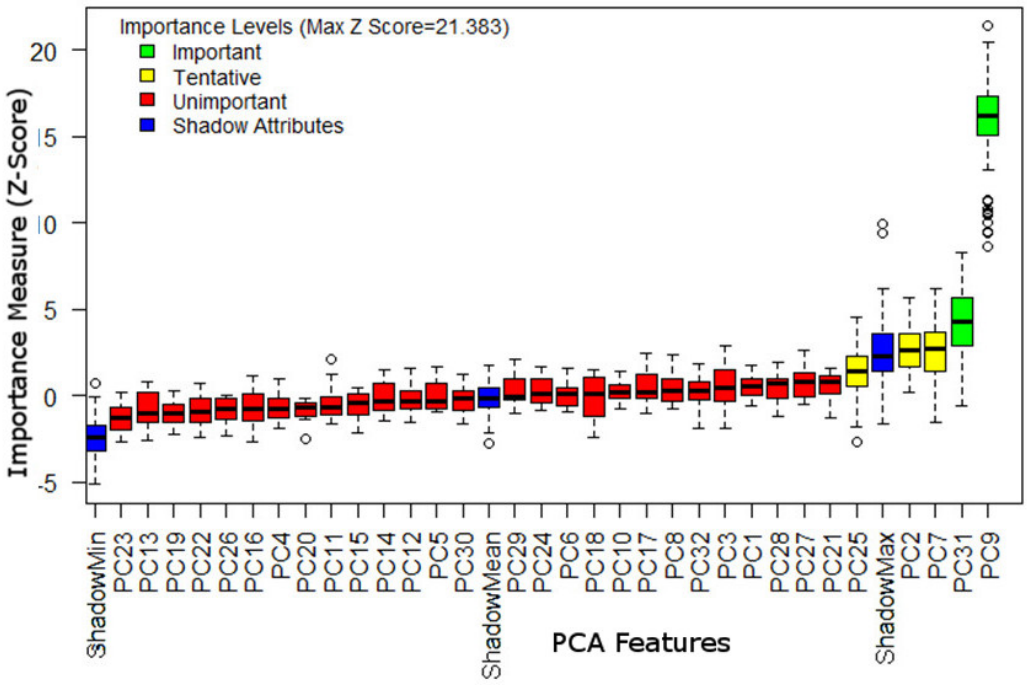



**Figure 5.10** Principal Component Importance levels for all features

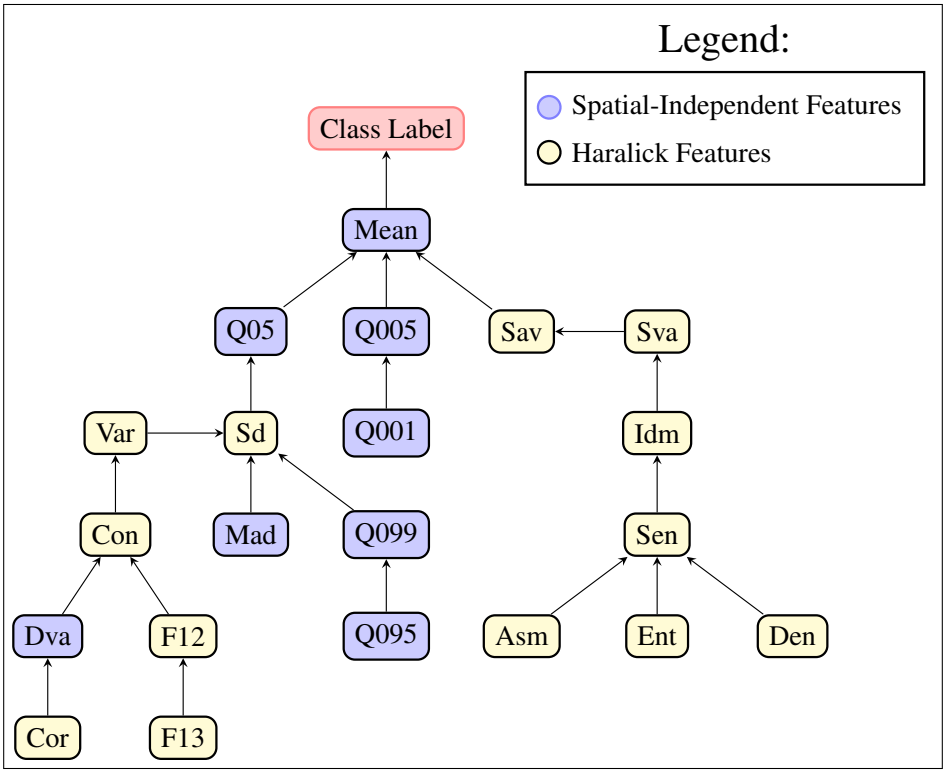



**Figure 5.11** Bayesian network with Spatial-independent feature properties

**Table 5.2** Evaluation of feature preprocessing techniques fitted to classification models

| | PCA | Outliers | Boruta | PCA + Outliers | PCA + Boruta | Outliers + Boruta | Outliers + PCA + Boruta |
|---|---|---|---|---|---|---|---|
| **Random Forest** | 0.67 | 0.78 | 0.76 | 0.80 | 0.75 | 0.92 | 0.88 |
| **Linear Discrimant Analysis (LDA)** | 0.67 | 0.70 | 0.77 | 0.81 | 0.73 | 0.72 | 0.70 |
| **Logistic Model (GLM)** | 0.60 | 0.62 | 0.81 | 0.76 | 0.74 | 0.82 | 0.77 |
| **K-NN (k=1)** | 0.87 | 0.87 | 0.78 | 0.70 | 0.61 | 0.57 | 0.67 |
| **K-NN (k=10)** | 0.67 | 0.70 | 0.66 | 0.71 | 0.79 | 0.72 | 0.71 |
| **Naive Bayes with Gaussian** | 0.78 | 0.70 | 0.67 | 0.87 | 0.79 | 0.82 | 0.83 |

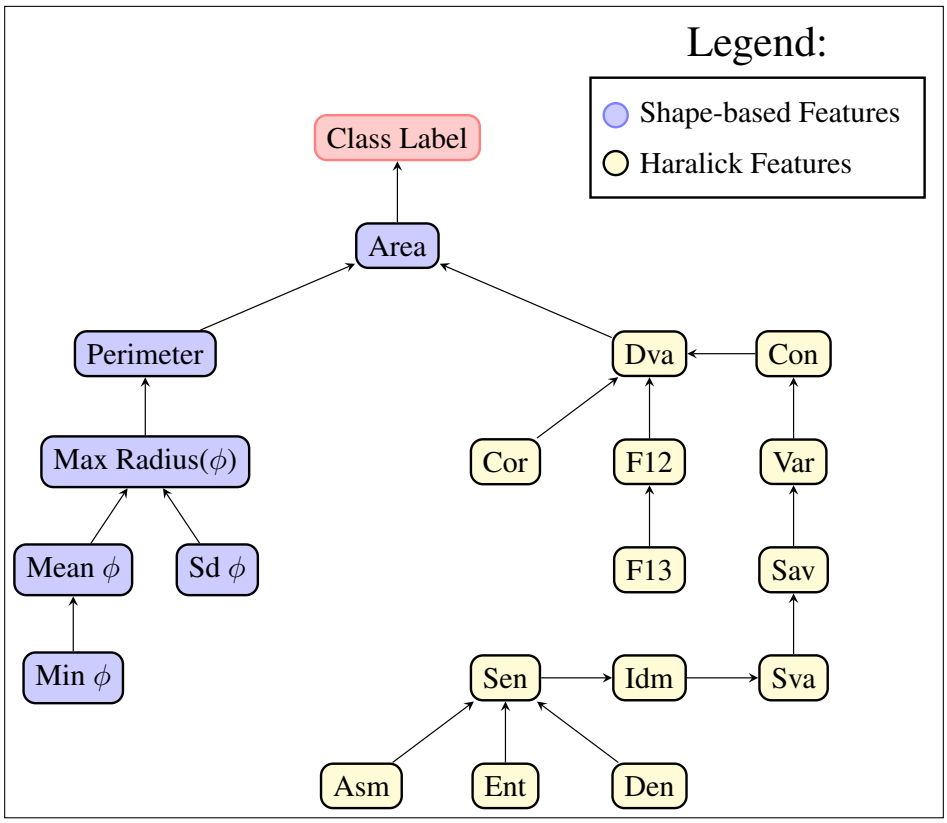



**Figure 5.12** Bayesian network with Shape feature properties

preprocessing techniques. The normality assumption is an important assumption for many statistical techniques and is illustrated in the dimensionality reductions of the topmost features with Kruskal measurements in Figure 5.9 and PCA components in Figure 5.10.

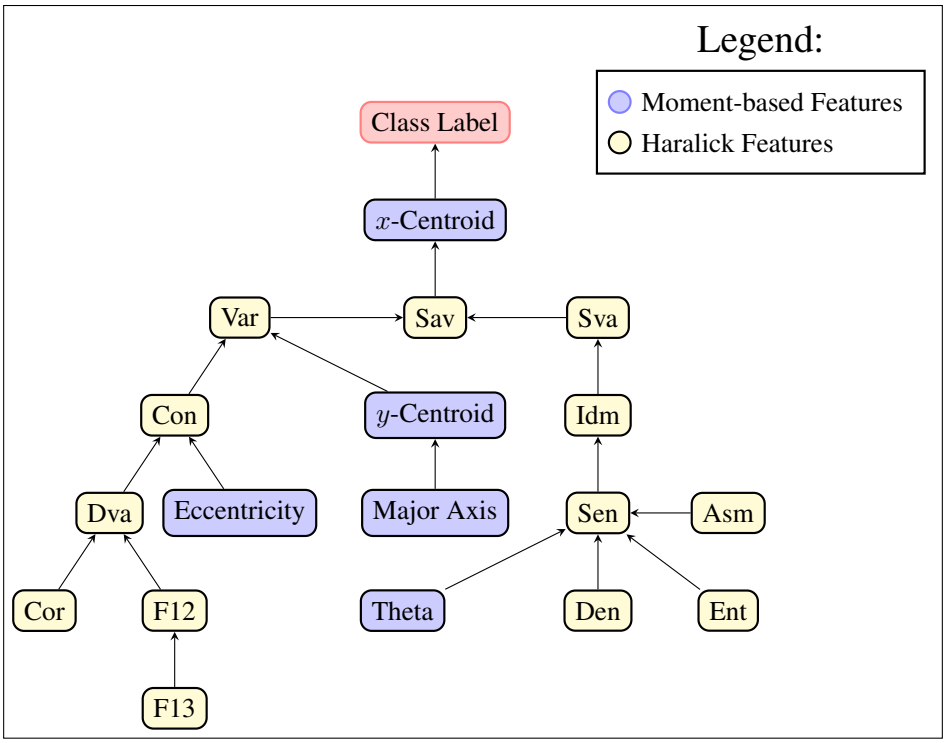



**Figure 5.13** Bayesian network of Moment feature properties

We later applied scaled entropy method by selecting a subset of the input variables with the highest information gain that influences the cancer label. The Information Gain $I_G(\cdot)$ in Equation (5.1) is the expected value of the entropy reduction of an instance $A$ attended from learning a set of random variables $S$.

$$I_G(S, A) = I_E(S) - \sum_{i=1}^{n} \left[ p\left(\subset_S^{A_i}\right) \times I_E\left(\subset_S^{A_i}\right) \right] \tag{5.1}$$

where;

- $I_E(S)$ and $I_E\left(\subset_S^{A_i}\right)$ is the Shannon Entropy of set $S$ and a subset $A_i$ of $S$ respectively, and
- $p\left(\subset_S^{A_i}\right)$ is the probability of the instance in a subset $A_i$ of $S \rightarrow \frac{size\left(\subset_S^{A_i}\right)}{size(S)}$.

## 5.4 Model Validation

We applied filtering of the data with filters kernels constructed from the proposed method and calculated gray-level co-occurrence matrices using the Haralick texture features. A subset of features from images with a significant height to-width ratio were then filtered and fitted to a classifier. A comparison of the proposed model with the BSIF, Gaussian, and Laplacian of Gaussian extraction methods was performed using a Random Forest and Recursive Partitioning classifiers on the training set.

The 40 split disjointed testing data set was then used to evaluate the proposed model and the classification performance was used to evaluate it. We fitted the independent testing data set to Recursive Partitioning and Random Forest Classifiers and evaluated their performance by measuring the classification Accuracy and Specificity (1-False-Positive Rate) against the Sensitivity (True-Positive Rate). The results of the experiments were evaluated by :

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5.3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

where $TP$ is the true positive, $TN$ is the true negative, $FP$ is the false positive and $FN$ is the false negative.

Thus, Specificity is 1 the proportion of the number of benign samples that were identified as malignant, while the Sensitivity is the proportion of malignant samples that are correctly identified as malignant.

### 5.4.1 Classifier Evaluation

The performance of a classifier can be evaluated by measuring it:

- *Accuracy*, it refers to how well a classifier predicts the correct class for unseen examples (that is, those not considered for learning the classifier);
- *Classification time*, how long it takes the classification process to predict the class, once the classifier has been trained;
- *Training time*, how much time is required to learn the classifier from data;
- *Memory requirements*, how much space in terms of memory is required to store the classifier parameters;
- *Clarity*, if the classifier is easily understood by a person.

### 5.4.2 Entropy Uncertainty

Entropy is a measure of uncertainties in an image. Let $f(x, y)$ be the image with various gray levels. The normalized histogram $C_i$ for an image of size $(A \times B)$ is given by:

$$C_i = \frac{P_i}{A \times B}, \quad \text{and} \tag{5.5}$$

$$\text{Shannon Entropy[63]:} S_e = -\sum_{i=0}^{R-1} C_i log_2 C_i \tag{5.6}$$

where $R$ represents the number of grey levels present in the given image.

$$\text{Renyi Entropy [63]:} R_e = \frac{1}{(1-\phi)} log_2 \sum_{i=0}^{R-1} C_i^{\phi} \tag{5.7}$$

where $\phi \neq 1, \phi > 0$.

$$\text{Kapur Entropy [21], [63]:} K_e = \frac{1}{\sigma - \phi} log_2 \frac{\sum_{i=0}^{R-1} C_i^{\phi}}{\sum_{i=0}^{R-1} C_i^{\sigma}} \tag{5.8}$$

where $\phi = \sigma, \phi > 0, \sigma > 0$.

For $\alpha \neq 1; \beta > 0; \alpha + \beta - 1 > 0$, where $\alpha$ and $\beta$ are diversity indices.

When $\alpha = 0.5$ and $\beta = 0.7$, we apply Yager's Measure as:

Yager's Measure [21], [63]:$Y_e$ $$=1 - \frac{\sum_{i=0}^{R-1} |2C_i - 1|)}{|A \times B|} \tag{5.9}$$

where $A$ and $B$ are rows and columns of the given image respectively.

## 5.5 Results

The analyses of the results obtained below show the classification performance from BSIF, Gaussian, and Laplacian filters against the proposed Adaptive Filter method in Figure 5.14. Before performing feature selection, we obtained an overall Accuracy of 75% with the Random Forest classifier and 83.33% with k-NN classifier where k=10 as illustrated in Table 5.3 and Table 5.4. After feature selection, we obtained an overall Accuracy of 92.88% with Random Forest and 86.21% with Recursive Partitioning. The Area Under the Curve in Figure 5.15.b was used to plot the Specificity and Sensitivity rates for the Random Forest classifier model for different filter sizes. A threshold is a percentage cut-off point that is used to make a decision when to classify a sample as positive (malignant). A low threshold value could increase the probability of a positive class detection (high sensitivity) but increase the false alarms (low specificity). Our model illustrates that a decision threshold of 0.99 will archive a 99.64% Sensitivity and 90.23% Specificity measure. We obtained the AUC to quantify the wellness of our model in distinguishing the two predictive values. The Elastic-Net parameters $\alpha$ = $0, 0.1, \cdots, 0.1$ and $\lambda$ = $0, 0.5 \times 10^2, \cdots, 10^3$ were used to generate kernel values for filter sizes = $(5 \times 5)$, $(7 \times 7)$, $\cdots$ ,$(17 \times 17)$. The AUC of 99.00% for the $(13 \times 13)$ Filter represents the probability for the model to correctly separate the benign cancer samples from malignant cancer.

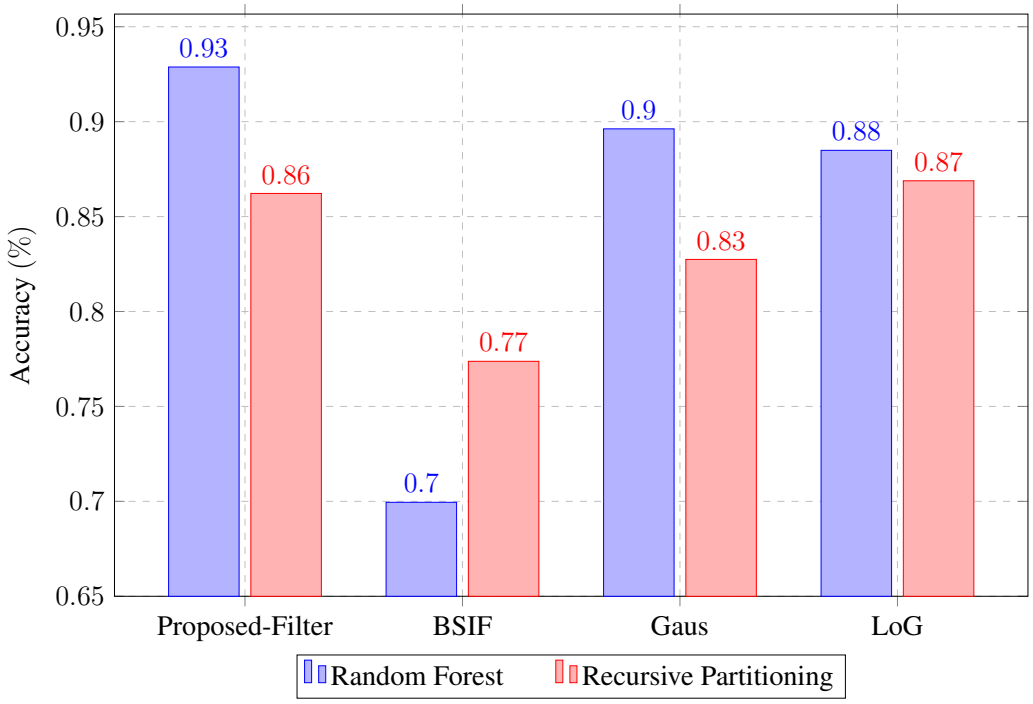


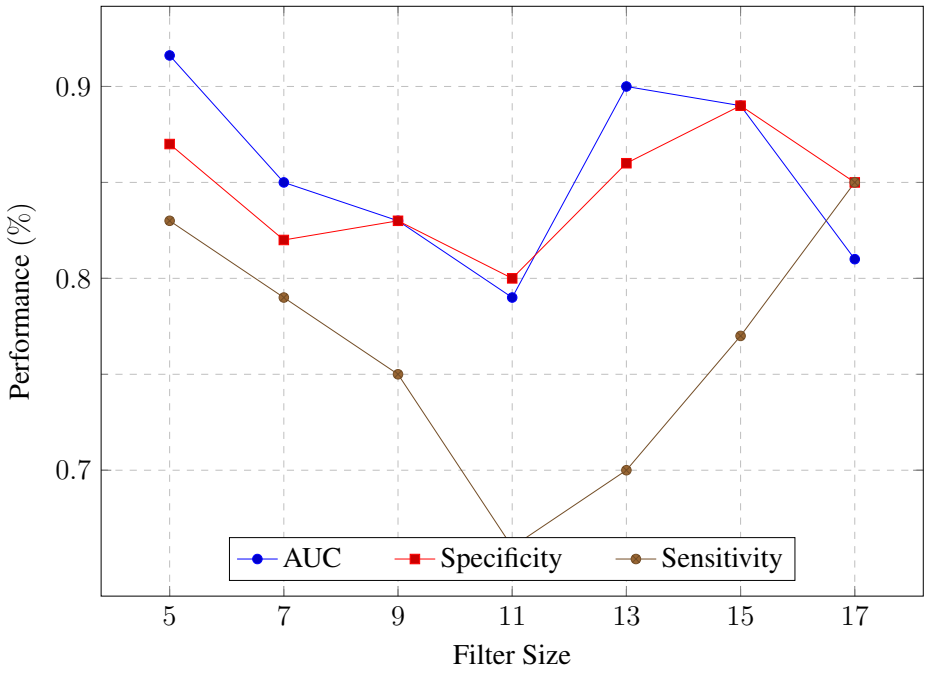

**Figure 5.14** Filter performances with the Random Forest and Recursive Partitioning classifiers

**Table 5.3** Evaluation of the proposed filter method fitted to various classifiers.

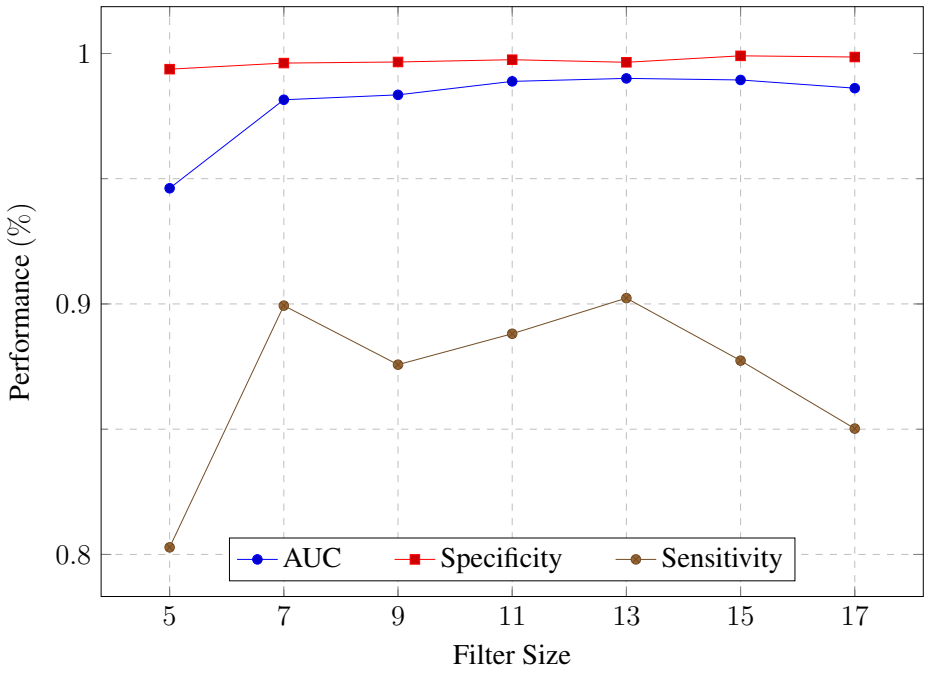
| **Extract:** | **Gaus** | | | **LoG** | | |
|---|---|---|---|---|---|---|
| **Features:** | **Haralick** | **Shape** | **Moment** | **Haralick** | **Shape** | **Moment** |
| **Random Forest** | 0.79 | 0.69 | 0.70 | 0.55 | 0.64 | 0.69 |
| **Recursive Partitioning** | 0.84 | 0.80 | 0.75 | 0.80 | 0.59 | 0.50 |
| **Naive Bayes** | 0.87 | 0.79 | 0.87 | 0.58 | 0.75 | 0.77 |
| **Logistic (GLM)** | 0.75 | 0.69 | 0.87 | 0.77 | 0.65 | 0.72 |
| **Support Vector Machine** | 0.48 | 0.47 | 0.42 | 0.57 | 0.61 | 0.68 |
| **k-NN (k = 10)** | 0.59 | 0.69 | 0.60 | 0.70 | 0.82 | 0.70 |

## 5.6 Discussion

Table 5.5 and Table 5.6 shows a summary of different CAD algorithms which researchers have employed in the study of distinguishing the benign from malignant thyroid




**(a)** Predictive performance of Random Forest classifier before feature selection




**(b)** Predictive performance of Random Forest classifier after feature selection

**Figure 5.15** Analysis plots above show the performance of the proposed filter method at different filter sizes Figure 5.15.a before and Figure 5.15.b after feature selection.

**Table 5.4** Evaluation of the proposed filter method fitted to various classifiers (continued).

| Extract: | BSIF | | | Proposed Method | | |
|---|---|---|---|---|---|---|
| **Features:** | **Haralick** | **Shape** | **Moment** | **Haralick** | **Shape** | **Moment** |
| **Random Forest** | 0.87 | *0.90 | 0.81 | 0.88 | 0.67 | 0.70 |
| **Recursive Partitioning** | 0.84 | 0.84 | 0.87 | 0.84 | 0.75 | 0.41 |
| **Naive Bayes** | 0.89 | 0.78 | 0.67 | 0.97 | 0.71 | 0.77 |
| **Logistic (GLM)** | 0.65 | 0.84 | 0.74 | 0.87 | 0.81 | 0.78 |
| **Support Vector Machine** | 0.67 | 0.42 | 0.47 | 0.68 | 0.72 | 0.58 |
| **k-NN (k = 10)** | 0.68 | 0.75 | 0.64 | *0.96 | 0.68 | 0.86 |

lesions with ultrasound images. Most of the researchers extracted highly significant textural features [20], [64]–[67] for the classification of the two classes (benign and malignant). Statistical [66] and higher-order statistical [68] based features are also extracted from the thyroid images. Discrete wavelet transform (DWT) is also a technique employed by researchers to denoise the thyroid images before the extraction of features [64], [69].

Further, an integrated index, namely: the thyroid malignancy index [64], [65], [69], and the thyroid clinical risk index [67], was formulated for a quick differentiation of benign or malignant thyroid lesions using a range of numerical values. Also, lately, a Convolutional Neural Network (CNN) model is implemented to distinguish benign and malignant thyroid lesions [70]. CNN is one of the promising machine learning techniques which requires the least pre-processing with no features extraction or selection and classifier [71], [72],[73]. This model can self-learn the necessary representations for the classification of the two classes (benign and malignant) through the different layers in the CNN architecture [71]. However, Chi et al. [70] incorporated the RF classifier to classify thyroid lesions. They did not fully utilize the benefits of a CNN model.

**Table 5.5** Selected studies on the CAD system for automated diagnosis of thyroid lesions (benign and malignant) with ultrasound images.

| **Authors (Year)** | **Number of images** | **Techniques** | **Performance** |
|---|---|---|---|
| Ding et al. (2011) [66] | Benign: 69<br>Malignant: 56 | Statistical and textural features, support vector machine classifier | Acc = 93.60%<br>Sen = 94.60%<br>Spec = 92.80% |
| Acharya et al. (2011) [69] | Benign: 400<br>Malignant: 400 | DWT, KNN classifier, thyroid malignancy index | Acc = 98.90%<br>Sen = 98.00%<br>Spec = 99.80% |
| Acharya et al. (2012a) [65] | HRUS<br>Benign: 400<br>Malignant: 400<br>CEUS<br>Benign: 400<br>Malignant: 400 | Grayscale texture, thyroid malignancy index | HRUS<br>Acc = 100.00%<br><br>CEUS<br>Acc = 98.10% |
| Acharya et al. (2012b) [64] | Benign: 400<br>Malignant: 400 | DWT, texture features, perceptron classifier, thyroid malignancy index | Acc = 100.00%<br>Sen = 100.00%<br>Spec = 100.00% |
| Acharya et al. (2012c) [68] | Benign: 400<br>Malignant: 400 | Higher-order statistics based, fuzzy classifier | Acc = 99.10%<br>Sen = 99.80%<br>Spec = 98.50% |
| Acharya et al. (2016) [20] | Benign: 211<br>Malignant: 31 | Gabor transform, textural features, decision tree classifier | Acc = 94.30% |

**Table 5.6** Selected studies on the CAD system for automated diagnosis of thyroid lesions (benign and malignant) with ultrasound images (continued).

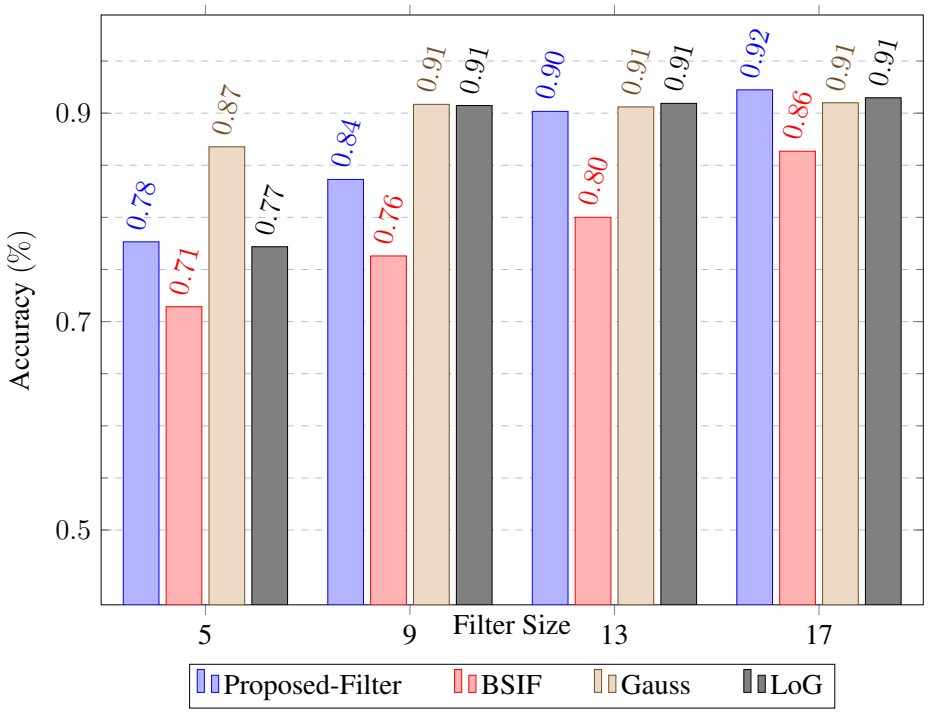
| **Authors (Year)** | **Number of images** | **Techniques** | **Performance** |
|---|---|---|---|
| Raghavendra et al. (2017) [67] | Benign: 211<br>Malignant: 31 | Spatial gray level dependence and fractal texture features, thyroid clinical risk index | Acc = 97.52%<br>Sen = 90.32%<br>Spec = 98.57% |
| Chi et al. (2017) [70] | **Public:**<br>Benign: 71<br>Malignant:357<br>**Private:**<br>Benign:129<br>Malignant: 35 | Convolutional neural network, random forest classifier | **Public:**<br>Acc = 98.29%<br>Sen = 99.10%<br>Spec = 93.90%<br>**Private:**<br>Acc = 96.34%<br>Sen = 86.00%<br>Spec = 99.00% |
| Present work | Benign: 246<br>Malignant: 142 | Adaptive Filter Method with Haralick Texture Features, Random Forest and Recursive Partitioning Classifiers | Acc = 96.00%<br>Sen = 99.64%<br>Spec = 90.23% |

Unlike traditional filtering methods that use heuristic methods of determining the filter parameters, we propose an adaptive method by fitting a Linear Regression Model on a subset of image pixel data and constructing the filter kernel from the coefficients estimated from the fitted model.

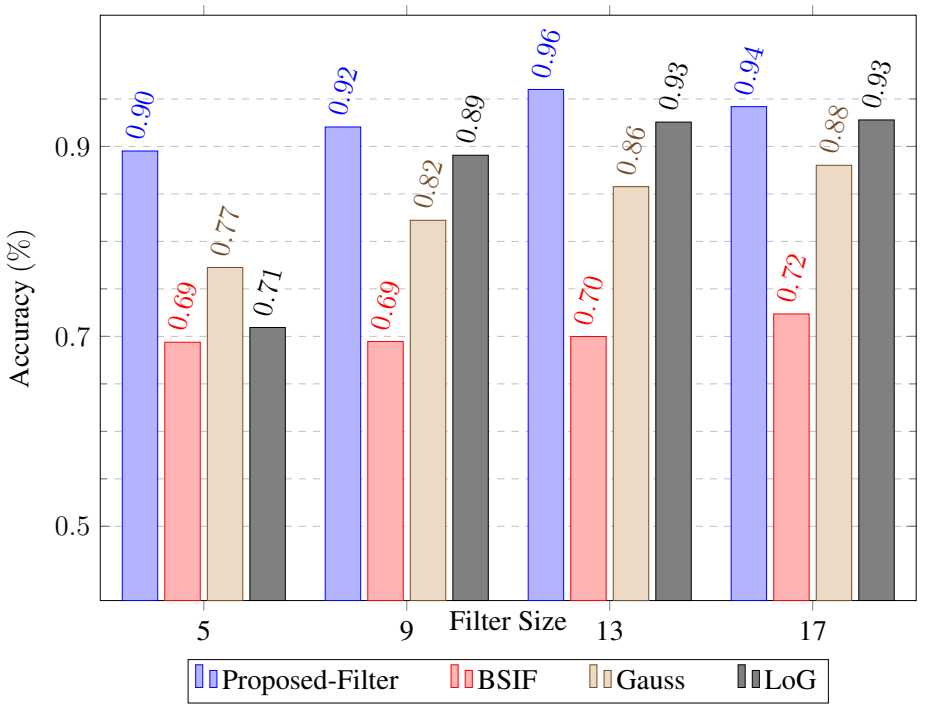
From the best of the authors' knowledge, no study has been done on the differentiation of benign and malignant thyroid lesions with the CNN. Nevertheless, recently Ma et al. [74] adopted the CNN model to automatically identify thyroid nodules in ultrasound

images. They achieved a high detection performance of 98.51%. However, their proposed CNN architecture could only pick up thyroid nodules from normal thyroid nodules. They did not extend to further categorize their thyroid nodules into either benign or malignant class. Therefore, the authors intend to adopt the CNN model to categorize the thyroid lesions into benign or malignant class in future studies. The CNN model is not employed in this work mainly due to the lack of images for both the benign and malignant thyroid lesions. Big data is required to yield maximum performance with CNN model.

In this work, a novel feature engineering model for extracting texture descriptors to differentiate the benign and malignant thyroid nodule. A high performance of 96.00% accuracy, with a sensitivity and specificity of 99.64% and 90.23% respectively (see Figure 5.16), is reported for filter size $13 \times 13$ in the differentiation of benign and malignant thyroid using the proposed framework. Nevertheless, the proposed methodology can be further improved for maximum detection accuracy and then be considered for implementation in the hospitals. This CAD system can be easily installed in the hospitals as it is portable and cost-effective. Also, hospitals can organize a mass screening for thyroid nodules annually for an early detection so that the thyroid nodules can be treated at an earlier stage. This can help to decrease the number of end-stage thyroid cancers in the society.




**(a)** Recursive Partitioning Classifier




**(b)** Random Forest Classifier

**Figure 5.16** Model Performance of (Figure 5.16.a) Recursive Partitioning and (Figure 5.16.b) Random Forest Classifiers using Image Filters from size 5 to 17.

# CHAPTER 6

# CONCLUSIONS

In this work, a novel feature engineering model for extracting texture descriptors to differentiate the benign and malignant thyroid nodule. A high performance of 96.00% accuracy, with a sensitivity and specificity of 99.64%) and 90.23% respectively, is reported for filter size 13 $\times$ 13 in the differentiation of benign and malignant thyroid using the proposed framework. Nevertheless, the proposed methodology can be further improved for maximum detection accuracy and then be considered for implementation in the hospitals. This CAD system can be easily installed in the hospitals as it is portable and cost-effective. Also, hospitals can organize a mass screening for thyroid nodules annually for an early detection so that the thyroid nodules can be treated at an earlier stage. This can help to decrease the number of end-stage thyroid cancers in the society.

## 6.1 Conclusions

It was observed that there was a boost in accuracies using the data-adaptive partitioning method to generate the membership function. To validate this finding, we compared our results with traditional non-fuzzy rule-based classification algorithms such as ANN, SVM, Decision trees, Naïve bayes, KNN, Rule induction, and ID3 (As shown in table 4). We know that these classification techniques work on the assumption that

all the features are independent of each other and the data is normally distributed. The Independent Component Analysis (ICA) was used to obtain a class-conditional representation to support this assumption. In ICA, multi-dimensional data is decomposed into components that are maximally independent in the Negentropy sense, uncovering disjoint underlying trends in the data.

### 6.1.1 Model Complexity

The direct computation of posterior probabilities in Bayesian Theorem is computationally expensive and an analysis of the complexity of a Bayesian Model is crucial for the successful application in a Bayesian Rule-based Decision System. The complexity of a simple problem can be determined by its ability to complete, thus deterministic or non-deterministic. A deterministic problem, e.g. One-dimensional State Machine, has a finite number of states and a known action to accept or halt a transition from one state to another. The time complexity of such a task is determined by the number of steps needed to complete it. This time is also not as polynomial time (P).

On the contrary, a non-deterministic problem requires a decision-making process to determine the transition action to perform, i.e., accept, reject or halt a task. The time complexity of a non-deterministic problem, referred to non-deterministic polynomial time (NP), is determined by the maximum possible steps the problem to satisfactory complete, given a defined set of input.

Other commonly referred complexity classifications for computational problems are the NP-Hard and NP-Complete. NP-hard are problems where there exists an algorithm for every NP problem that can reduce it in a polynomial time by a deterministic machine

while an NP-complete problems are the intersection of NP and NP-Hard problems. Note that all NP-Complete problems are NP Hard, but not all NP Hard problems are NP-complete. Bayesian Network inference problems are generally NP-Hard that can be

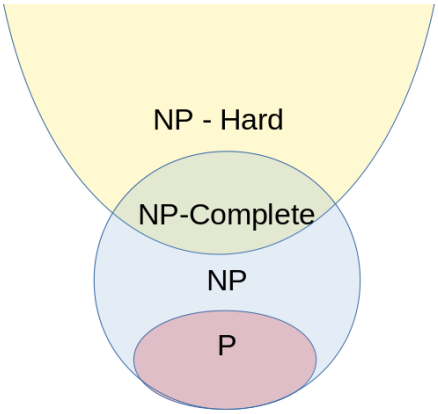



**Figure 6.1** The relationship between complexity of different classes

reduced in polynomial time by identifying the optimal Directed Acyclic Graph (DAG) by using techniques like the Markov Chain Monte Carlo (MCMC) methods [53], Metropolis-Hastings algorithm [61] and Gibbs Sampling[75].

## 6.2 Future Work

The work presented in this dissertation has is based on the description of natural phenomena in the research of biomedical image diagnosis research and we recommend further exploration in this study in the future. We identified that data mining techniques for extracting domain relevant texture information from natural images are not domain specific to the area of application. This opens up avenues for research into domain-adaptive techniques as an alternative to the data-adaptive technique introduced in this study. Secondly, the recent improvement in technology (both in software and hardware) has made it possible to design more efficient computational models that are cost and resource

effective. However, we noted that thyroid ultrasound image diagnostic techniques are not as comprehensive as other image diagnostic techniques that use MRI and X-Ray due to the limitation in thyroid diagnosis process noted in Figure 1.2. We envision research in compatibility of traditional data mining models like the Bayesian Inference to Hyper Computing Technologies like the used of Graphics Processing Unit (GPU).

# APPENDIX A

# REGULARIZATION

**Algorithm 1:** Ridge regression algorithm

**Input :** Independent variables : $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
Target variable: $\{y_1, \ldots, y_m\}$

**Output:** Model predictions $F(*) = \{F(\mathbf{x}_1), \ldots, F(\mathbf{x}_m)\}$

**Function** `Ridge`$(Y, X, \gamma)$ **:**
    Initialize $\mathbf{A} = \gamma\mathbb{I}$;
    $\mathbf{b} = 0$;
    **For** $t = 1$ *to* $m$ **Do:**
        Read new $\mathbf{x}_t$;
        $F(\mathbf{x}_t) = \mathbf{b}^\intercal \mathbf{A}^{-1}\mathbf{x}_t$ (prediction) ;
        $\mathbf{A} = \mathbf{A} + \mathbf{x}_t\mathbf{x}_t^\intercal$;
        Read new $y_t$;
        $\mathbf{b} = \mathbf{b} + y_t\mathbf{x}_t$;
    **End**
**return** $F(*)$

# APPENDIX B

# CROSS VALIDATION

**Algorithm 2:** Cross validation procedure using k folds

```
Input  : DataSet, k number of folds
Output: CVModels
Function CrossValidation(DataSet, k):
    // Initialize Parameters:
    // Randomly split dataset into k disjoint sets
    SubSet[1, ··· , k] ← DataSet ;
    For i = 1 → k Do:
        TrainSet ← SubSet[1, ··· , i − 1, ··· , k] ;
        TestSet ← SubSet[i] ;
            TrainModel ← Train(TrainSet) ;
            Classifier ← Classify(TrainModel, TestSet) ;
            Predict ← ConfusionMatrix(Classifier) ;
        Append(CVModels, Score(Predict));
    End
return CVModels
```

# APPENDIX C

# IMAGE FILTERING WITH THE DISCRETE FOURIER TRANSFORM (DFT) ALGORITHM

**Algorithm 3:** Discrete fourier transform (DFT) algorithm

**Input :** Data, flags
**Output:** DFTData

**Function** DFTMTX($N$)**:**
  **return** $\mathbf{W}_{x,y} = e^{\frac{j2\pi xy}{N}}$;
**Function** IDFTMTX($N$)**:**
  **return** $\mathbf{W_i}_{x,y} = e^{\frac{-j2\pi xy}{N}}$;
**Program** DFT2D(Data, *flags*)**:**
  $M$ = number of rows in Data;
  $N$ = number of columns in Data;
  **if** *flags = 1* **then**
    $\mathbf{W_M}$ = DFTMTX($M$)
    $\mathbf{W_N}$ = DFTMTX($N$)
    **return** $\mathbf{W_M f W_N}$;
  **else**
    $\mathbf{W_M^i}$ = IDFTMTX($M$);
    $\mathbf{W_N^i}$ = IDFTMTX($N$);
    **return** $\frac{1}{MN}\mathbf{W_M^i f^i W_N}$;
  **end**
**return** DFTData

# APPENDIX D

# FAST FOURIER TRANSFORM

**Algorithm 4:** 2-D fast fourier transform algorithm

**Input :** $A$
**Output:** $\hat{A}$

**Function** `FFT(`$k, \omega$`)`**:**
  Initialize: $\hat{A}(x)$, $\hat{A}(x^E)$, $\hat{A}(x^O)$;
  RESIZE($A(x), 2^k$)
  **if** $k == 0$ **then**
    $\hat{A}(x) = A(x_0)$;
  **else**
    $A(x^E) = A(x_0), A(x_2), \cdots, A(x_{2^{k-2}})$ ;
    $A(x^O) = A(x_1), A(x_3), \cdots, A(x_{2^{k-1}})$ ;
    $\hat{A}(x^E)$ = FFT($A(x^E), k-1, \omega^2$) ;
    $\hat{A}(x^O)$ = FFT($A(x^O), k-1, \omega^2$) ;
  **end**
  **forall** $i = 0 \rightarrow 2^{k-1}$ **do**
    $\hat{A}(x_i) = \hat{A}(x_i^E) + \omega^i \cdot \hat{A}(x_i^O)$ ;
    $\hat{A}(x_{i+2^{k-1}}) = \hat{A}(x_i^E) - \omega^i \cdot \hat{A}(x_i^O)$ ;
  **end**
**return** $\hat{A}$

# APPENDIX E

# CONVOLUTION IMAGE FILTER ALGORITHM

**Algorithm 5:** Image filter algorithm

**input :** Original Image $Orig\_img(w \times h)$, Filter $Filt\_img(u \times v)$
**output:** Processes Image $Conv\_img$

```
1  Function Convolution(Orig_img, Filt_img):
2      Conv_img ← new: Image(w × h) ;
3      for y = 0 to h do
4          for x = 0 to w do
5              sum ← 0;
6              for j = -v to v do
7                  for i = -u to u do
8                      sum ← sum + Filt_img(i, j) × Orig_img(x + i, y + j);
9                  end
10             end
11             Conv_img(x, y) = sum;
12         end
13     end
14 return Conv_img
```

# APPENDIX F

# ALGORITHM FOR EXTRACTING HARALICK IMAGE FEATURES

**Algorithm 6:** Haralick image feature algorithm

**input :** $Img$ Input image (width $w$, height $h$), $\rho_0$ threshold condition
**output:** $Img_{OUT}$

**Function** `Haralick(`$Img, Filt\_img$`)`**:**
  `// 5 × 5 masks is used to estimate coefficients` $k_1 \cdots k_{25}$
  $im \leftarrow$ input image;
  **forall** *pixel* $i$ *in image* $im$ **do**
    $neighbors \leftarrow 5 \times 5$ neighborhood of pixel $i$ in image $im$;
    **for** $j = 1 \cdots 10$ **do**
      $k_j \leftarrow$ `Convolution(`$neighbors$,$mask[j]$`)`;
    **end**
    $C_2 \leftarrow \frac{k_2^2 k_4 + k_2 k_3 k_5 + k_3^2 k_6}{k_2^2 + k_3^2}, C_3 \leftarrow \frac{k_2^3 k_7 + k_2^2 k_3 k_8 + k_2 k_3^2 k_9 + k_3^3 k_{10}}{(\sqrt{k_2^2 + k_3^2})^3}$;
    **if** $|C_2/3C_3| \leq \rho_0$ & $C_3 < 0$ **then**
      $Img_{OUT}[i] \leftarrow 255$;
    **else**
      $Img_{OUT}[i] \leftarrow 0$;
    **end**
  **end**
**return** $Img_{OUT}$

# APPENDIX G

# BAYESIAN APPROACH

**Table G-1** Table of symbols used in this study.

| Notation | Description |
|---|---|
| Main Dataset | S |
| Training Dataset | $D_{train}$ |
| Testing Dataset | $D_{test}$ |
| Number of points of the data set $S$ | $N_s$ |
| Number of points of the data set $D_{trian}$ | $N_{tr}$ |
| Number of points of the data set $D_{test}$ | $N_{test}$ |
| One-dimensional Target Class Variable of size $n$ Rows $\times$ 1 Column | $Y$ |
| Two-dimensional Input Variable of size $n$ Rows $\times$ $p$ Columns | $X$ |
| Values corresponding to the Variables $X$ and $Y$ | $x, y$ |
| Input values at Column $i$ | $x_{I,*}$ |
| Input values at Row $j$ | $x_{*,j}$ |
| Set of edges of the $BN$ | $E$ |
| Set of parameters of local pdfs for entire $BN$ | $T$ |
| Number of edges of the $BN$ | $m$ |
| Directed Acyclic Graph (DAG) of a $BN$ | $G$ |
| Markov blanket of variable $X$ | $B(X)$ |
| Set of direct neighbors of variable $X$ in Bayesian network | $N(X)$ |
| Set of parents of $x_{I,*}$ | $Pa_i$ |
| Set of parents $Pa_i$ of $x_{*,i}$ for each member assignment $j$ | $pa_{I,j}$ |
| Number of values of discrete variable $x_{*,i}$ | $r_i$ |
| Number of configurations of the set of parents of $x_{*i}$ | $q_i$ |
| Counts of a multinomial distribution with $K$ bins | $c_1$ |
| Parameters (bin probabilities) of a multinomial distribution | $p_1$ |
| Penalty Parameters of the Logistic Regression Function | $\lambda\alpha$ |
| Coefficients of the Logistic Regression Function | $\beta$ |

**Table G-2** Bayesian Theory notations

| **Notation** | **Expression** |
|---|---|
| Likelihood | $p(x\|\theta)$ |
| Prior Probability | $p(\theta)$ |
| Posterior Probability | $p(\theta\|x)$ |
| Expectation | $p(x)$ |
| Posterior Predictive | $p(x_{n+1}\|x_{1:n})$ |
| Loss Function / Posterior Expected Loss | $\ell(s, a)$ |

# LIST OF ABBREVIATIONS

**BMI** Biomedical Informatics. vii, 1–3

**BN** Bayesian Networks. ix, 46

**BNS** Bayesian Network Structure. x, 49, 63

**BSIF** Binarized Statistical Image Features. xi, 19, 20, 32, 33, 42, 67, 69

**CDSS** Clinical Decision Support Systems. vii, 3, 4, 12

**CNN** Convolutional Neural Network. 72, 74, 75

**CP** Conditional Probability. 47, 48

**CT** Computed Tomography. 5, 7

**Health-IT** Health Information Technology. 1, 2

**HRL** Haralick. viii, ix, xiii, 21–24, 30, 39, 42, 60, 63, 67, 74, 92

**JP** Joint Probability. 46, 47, 49

**MP** Marginal Probability. 47, 49

**MRI** Magnetic Resonance Image. 5–7

**US** Ultrasound. vii, 5–7, 11

# BIBLIOGRAPHY


[1] Antwerp University Hospital - Universitair Ziekenhuis Antwerpen (UZA).

[2] National Institutes of Health, Thyroid Cancer-Patient Version, en, 2017.

[3] M. I. Surks, E. Ortiz, G. H. Daniels, C. T. Sawin, N. F. Col, R. H. Cobin, J. A. Franklyn, J. M. Hershman, K. D. Burman, M. A. Denke, C. Gorman, R. S. Cooper, and N. J. Weissman, Subclinical Thyroid Disease: Scientific Review and Guidelines for Diagnosis and Management, 2004.

[4] E. V. Bernstam, J. W. Smith, and T. R. Johnson, "What is biomedical informatics?" en, Journal of Biomedical Informatics, vol. 43, no. 1, pp. 104–110, 2010.

[5] J. R. Staniland, J. C. Horrocks, and F. T. De Dombal, "Computer—assisted Diagnosis of Abdominal Pain Using "Estimates" Provided by Clinicians," British Medical Journal, 1972.

[6] K. Doi, Diagnostic imaging over the last 50 years: Research and development in medical imaging science and technology, 2006.

[7] M Henderson and J Dolan, "Challenges, solutions, and advances in ultrasound-guided regional anaesthesia," en, BJA Education, vol. 16, no. 11, pp. 374–380, 2016.

[8] H. Gharib, E. Papini, R. Paschke, D. S. Duick, R. Valcavi, L. Hegedüs, and P. Vitti, "American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules," Journal of Endocrinological Investigation, vol. 33, no. 5 SUPPL. Pp. 1–50, 2010.

[9] J. Norman and G. Clayman, Thyroid Nodules: Hyperthyroidism and Thyroid Cancer, en, 2017.

[10] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," CA: A Cancer Journal for Clinicians, vol. 67, no. 1, pp. 7–30, 2017.

[11] American Thyroid Association, Thyroid Nodules, 2017.

[12] V. Chan and A. Perlas, "Basics of ultrasound imaging," in Atlas of Ultrasound-Guided Procedures in Interventional Pain Management, Springer, 2011, pp. 13–19.

[13] N. Larburu, R. Bults, M. Van Sinderen, and H. Hermens, "Quality-of-data management for telemedicine systems," in Procedia Computer Science, 2015.

[14] R. C. Gonzalez and R. E. Woods, "Digital Image Processing, Fourth Edition," Pearson Education International, 2018.

[15] S. Ibrahim, P. Chowriappa, S. Dua, U. R. Acharya, K. Noronha, S. Bhandary, and H. Mugasa, "Classification of diabetes maculopathy images using data-adaptive neuro-fuzzy inference classifier," Medical and Biological Engineering and Computing, vol. 53, no. 12, pp. 1345–1360, 2015.

[16] T. K. Boehme and R. Bracewell, "The Fourier Transform and its Applications.," The American Mathematical Monthly, 2006.

[17] J Kannala and E Rahtu, "BSIF: Binarized statistical image features," in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 1363–1366.

[18] M. Jain and P. Dua, "Data Adaptive Rule-based Classification System for Alzheimer Classification," Journal of Computer Science & Systems Biology, 2013.

[19] M. R. K. Mookiah, U. Rajendra Acharya, C. M. Lim, A. Petznick, and J. S. Suri, "Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features," Knowledge-Based Systems, 2012.

[20] U. R. Acharya, P. Chowriappa, H. Fujita, S. Bhat, S. Dua, J. E. Koh, L. W. Eugene, P. Kongmebhol, and K. H. Ng, "Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images," Knowledge-Based Systems, vol. 107, pp. 235–245, 2016.

[21] U. Rajendra Acharya, M. R. K. Mookiah, S. Vinitha Sree, R. Yanti, R. Martis, L. Saba, F. Molinari, S. Guerriero, and J. S. Suri, "Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification," in Ovarian Neoplasm Imaging, 2013.

[22] A. Sehad, Y. Chibani, R. Hedjam, and M. Cheriet, "Gabor filter-based texture for ancient degraded document image binarization," Pattern Analysis and Applications, vol. 22, no. 1, pp. 1–22, 2019.

[23] R. M. Haralick, I. Dinstein, and K. Shanmugam, "Textural Features for Image Classification," IEEE Transactions on Systems, Man and Cybernetics, 1973.

[24] D. Sudheer, R. SethuMadhavi, and P. Balakrishnan, "Edge and Texture Feature Extraction Using Canny and Haralick Textures on SPARK Cluster," in Proceedings of the 2nd International Conference on Data Engineering and Communication Technology. Springer, 2019.

[25] I. Vrbik, S. J. Van Nest, P. Meksiarun, J. Loeppky, A. Brolo, J. J. Lum, and A. Jirasek, "Haralick texture feature analysis for quantifying radiation response heterogeneity in murine models observed using Raman spectroscopic mapping," PLoS ONE, vol. 14, e0212225, Feb. 2019.

[26] X. Du, S. Dua, R. U. Acharya, and C. K. Chua, "Classification of Epilepsy Using High-Order Spectra Features and Principle Component Analysis," Journal of Medical Systems, vol. 36, no. 3, pp. 1731–1743, 2012.

[27] H. Ishibuchi, Y. Kaisho, and Y. Nojima, "Designing fuzzy rule-based classifiers that can visually explain their classification results to human users," in 2008 3rd International Workshop on Genetic and Evolving Fuzzy Systems, GEFS, 2008.

[28] S. Dua and S. Saini, "Data Shrinking Based Feature Ranking for Protein Classification," in Information Systems, Technology and Management. Springer, 2009.

[29] H. Singh, P. Chowriappa, and S. Dua, "Multi-domain Protein Family Classification Using Isomorphic Inter-property Relationships," in Contemporary Computing. Springer, 2009.

[30] J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," IEEE Transactions on Systems, Man and Cybernetics, 1993.

[31] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Y. Ng, and A. Laude, Computer-aided diagnosis of diabetic retinopathy: A review, 2013.

[32] J. Nayak, P. S. Bhat, R. Acharya U, C. M. Lim, and M. Kagathi, "Automated identification of diabetic retinopathy stages using digital fundus images," Journal of Medical Systems, 2008.

[33] P. Chowriappa, S. Dua, U. Rajendra Acharya, and M. Muthu Rama Krishnan, "Ensemble selection for feature-based classification of diabetic maculopathy images," Computers in Biology and Medicine, 2013.

[34] S. Sahu, H. V. Singh, B. Kumar, and A. K. Singh, "De-noising of ultrasound image using Bayesian approached heavy-tailed Cauchy distribution," Multimedia Tools and Applications, vol. 78, no. 4, pp. 4089–4106, 2019.

[35] Y. Nagaraj and A. V. Narasimhadhan, "Comparison of Edge Detection Algorithms in the Framework of Despeckling Carotid Ultrasound Images Based on Bayesian Estimation Approach," in Computer Vision, Pattern Recognition, Image Processing, and Graphics. Springer, 2018.

[36] S. Dua, U. Rajendra Acharya, P. Chowriappa, and S. Vinitha Sree, "Wavelet-based energy features for glaucomatous image classification," IEEE Transactions on Information Technology in Biomedicine, 2012.

[37] U. R. Acharya, S. Dua, X. Du, V. Sree S, and C. K. Chua, "Automated diagnosis of glaucoma using texture and higher order spectra features," IEEE Transactions on Information Technology in Biomedicine, 2011.

[38] S. M. Camps, D. Fontanarosa, P. H. N. de With, F. Verhaegen, and B. G. L. Vanneste, "The Use of Ultrasound Imaging in the External Beam Radiotherapy Workflow of Prostate Cancer Patients," BioMed Research International, vol. 2018, pp. 1–16, 2018.

[39] S. Mallat, A Wavelet Tour of Signal Processing [3rd Edition], en, Second Edi, 3. Elsevier, 2009, vol. 11, p. 620.

[40] M. Nieniewskib and P. Zajaczkowski, Comparison of ultrasound image filtering methods by means of multivariable kurtosis, en, 2017.

[41] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations.," Computer vision, graphics, and image processing, vol. 39, no. 3, pp. 355–368, 1987.

[42] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014, 2014, pp. 2392–2397.

[43] X. Wang, B. S. Wong, and T. C. Guan, "Image enhancement for radiography inspection," in SPIE Proceedings vol. 5852, Third International Conference on Experimental Mechanics and Third Conference of the Asian Committee on Experimental Mechanic, 2005, pp. 462–468.

[44] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[45] M. Zaffar, M. A. Hashmani, and K. S. Savita, "Comparing the Performance of FCBF, Chi-Square and Relief-F Filter Feature Selection Algorithms in Educational Data Mining," in Recent Trends in Data Science and Soft Computing. Springer, 2019.

[46] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain," in Advances in Artificial Intelligence - SBIA 2012. Springer, 2012.

[47] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol. 46, no. 1-3, pp. 389–422, 2002.

[48] G Yun and S Reeves, “Efficient Computation for Sequential Forward Observation Selection in Image Reconstruction,” 1998.

[49] S. Gunasundari and S. Janakiraman, “A Hybrid PSO-SFS-SBS Algorithm in Feature Selection for Liver Cancer Data,” in Power Electronics and Renewable Energy Systems. Springer, 2015.

[50] W. R. R. Miron B. Kursa, M. B. Kursa, W. R. Rudnicki, and Others, “Feature Selection with the Boruta Package,” Journal Of Statistical Software, vol. 36, no. 11, pp. 1–13, 2010.

[51] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, “A Survey on Deep Learning in Medical Image Analysis,” arXiv:1702.05747 [cs], 2017.

[52] D. MacKenzie and S. M. Stigler, The History of Statistics: The Measurement of Uncertainty before 1900, 2. 2006, vol. 29, p. 299.

[53] G. Ridgeway and D. Madigan, “A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets,” in Data Mining and Knowledge Discovery, 2003.

[54] Dimitris Margaritis, “Learning Bayesian Network Model Structure from Data,” PhD thesis, 2003.

[55] I. Tsamardinos, C. Aliferis, A. Statnikov, and E Statnikov, “Algorithms for Large Scale Markov Blanket Discovery.,” FLAIRS Conference, 2003.

[56] S. Yaramakala and D. Margaritis, “Speculative Markov blanket discovery for optimal feature selection,” in Proceedings - IEEE International Conference on Data Mining, ICDM, 2005.

[57] D. Heckerman, “A tutorial on learning with Bayesian networks,” Studies in Computational Intelligence, 2008.

[58] A. Delaplace, T. Brouard, and H. Cardot, “Two evolutionary methods for learning Bayesian network structures,” in 2006 International Conference on Computational Intelligence and Security, ICCIAS 2006, 2007.

[59] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, “The max-min hill-climbing Bayesian network structure learning algorithm,” Machine Learning, 2006.

[60] M. Gasse, A. Aussem, and H. Elghazel, “A hybrid algorithm for Bayesian network structure learning with application to multi-label learning,” Expert Systems with Applications, 2014.

[61] C. P. Robert and G. Casella, “Metropolis–Hastings Algorithms,” in Introducing Monte Carlo Methods with R. Springer, 2010.

[62] Universidad Nacional de Colombia — Computer Imaging and Medical Applications Laboratory.

[63] A. P. S. Pharwaha and B. Singh, "Shannon and Non-Shannon Measures of Entropy for Statistical Texture Feature Extraction in Digitized Mammograms," in Wcecs 2009: World Congress on Engineering and Computer Science, Vols I and Ii, 2009.

[64] U. R. Acharya, O. Faust, S. V. Sree, F. Molinari, and J. S. Suri, "ThyroScreen system: High resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform," Computer Methods and Programs in Biomedicine, vol. 107, no. 2, pp. 233–241, 2012.

[65] U. R. Acharya, S. V. Sree, M. R. K. Mookiah, F. Molinari, R. Garberoglio, and J. S. Suri, "Non-invasive automated 3D thyroid lesion classification in ultrasound: A class of ThyroScan systems," Ultrasonics, vol. 52, no. 4, pp. 508–520, 2012.

[66] J. Ding, H. Cheng, C. Ning, J. Huang, and Y. Zhang, "Quantitative measurement for thyroid cancer characterization based on elastography," Journal of Ultrasound in Medicine, vol. 30, no. 9, pp. 1259–1266, 2011.

[67] U. Raghavendra, U. Rajendra Acharya, A. Gudigar, J. Hong Tan, H. Fujita, Y. Hagiwara, F. Molinari, P. Kongmebhol, and K. Hoong Ng, "Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions," Ultrasonics, vol. 77, pp. 110–120, 2017.

[68] U. Acharya, S. Sree, G. Swapna, S. Gupta, F. Molinari, R. Garberoglio, A. Witkowska, and J. Suri, "Effect of complex wavelet transform filter on thyroid tumor classification in three-dimensional ultrasound," Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, vol. 227, no. 3, pp. 284–292, 2013.

[69] U. R. Acharya, O. Faust, V. S. Sree, F. Molinari, R. Garberoglio, and S. J. Suri, "Cost-Effective and Non-Invasive Automated Benign and Malignant Thyroid Lesion Classification in 3D Contrast-Enhanced Ultrasound Using Combination of Wavelets and Textures: A Class of ThyroScan(tm) Algorithms.," Technology in Cancer Research and Treatment, vol. 10, no. 4, pp. 371–380, 2011.

[70] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network," Journal of Digital Imaging, vol. 30, no. 4, pp. 477–486, 2017.

[71] Y. Lecun, Y. Bengio, and G. Hinton, Deep learning, 2015.

[72] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," Journal of Computational Science, vol. 20, pp. 70–79, 2017.

[73] J. H. Tan, H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao, K. C. Chua, and U. R. Acharya, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," Information Sciences, vol. 420, pp. 66–76, 2017.

[74] J. Ma, F. Wu, T. Jiang, J. Zhu, and D. Kong, "Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images," Medical Physics, vol. 44, no. 5, pp. 1678–1691, 2017.

[75] M. Hulme, Improved Sampling for Diagnostic Reasoning in Bayesian Networks, Feb. 2013.