Novel Statistical and Probabilistic Machine Learning Algorithms for Genotype Clustering and Cancer Classification

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

by

Aditya A. Shastri

1501201001

Supervisor:

Dr. Kapil Ahuja

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

INDORE - 453 552

2021

Novel Statistical and Probabilistic Machine Learning Algorithms for Genotype Clustering and Cancer Classification

By

Aditya A. Shastri

A Thesis Submitted to

Indian Institute of Technology Indore

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Approved:

_____

Dr. Kapil Ahuja

Thesis Advisor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

INDORE - 453 552

February 2021

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled "**Novel Statistical and Probabilistic Machine Learning Algorithms for Genotype Clustering and Cancer Classification**" in the partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in the **Discipline of Computer Science and Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from January 2016 to February 2021 under the supervision of Dr. Kapil Ahuja, Associate Professor, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

(10-02-2021)

Signature of the student with date

(Mr. ADITYA ANAND SHASTRI)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(10-02-2021)

Signature of Thesis Supervisor with date

(Dr. KAPIL AHUJA)

Mr. ADITYA ANAND SHASTRI has successfully given his Ph.D. Oral Examination held on 28 July 2021.

Signature of Chairperson (OEB)	Signature of External Examiner	Signature of Thesis Supervisor
Date: 28-07-2021	Date: 28-07-2021	Date: 28-07-2021
(Dr. Manavendra Mahato, IITI)	**(Prof. C Krishna Mohan, IITH)**	**(Dr. Kapil Ahuja, IITI)**

Signature of PSPC Member #1	Signature of PSPC Member #2
Date: 28-07-2021	Date: 28-07-2021
(Dr. Aruna Tiwari, IITI)	**(Dr. Trapti Jain, IITI)**

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude towards my supervisor **Dr. Kapil Ahuja** for his continuous support in the journey of my Ph.D. study and related research. His patience, motivation, guidance, and immense knowledge have helped me in all the time of research, writing research articles, and this dissertation. He timely encouraged me with his invaluable advice, positive criticisms and stimulating discussions that allowed me to grow as a researcher. He kindly corrected me at every stage of this journey with his expertise and brilliance in coming up with new ideas, implementing and presenting them with great transparency and writing precisely. Without his tremendous understanding, this learning process would have been impossible. I could not have imagined having a better adviser and mentor for my Ph.D. study.

Besides my supervisor, I want to offer my special thanks to **Dr. Milind Ratnaparkhe** for his assistance in better understanding the bioinformatics concepts. He was always available and approachable for the technical discussions, even at the eleventh hour. Furthermore, I want to express my sincere gratitude to **Prof. Yann Busnel**, who provided me an opportunity to join his team as an intern at IMT Atlantique, Rennes, France. I am also thankful to **Dr. Deepti Tamrakar** for her help and conversations regarding breast cancer research.

I would also like to thank the rest of my research progress committee members: **Dr. Aruna Tiwari** and **Dr. Trapti Jain**, for their insightful comments and encouragement. Their suggestions and questions have incented me to widen my research from various perspectives. Dr. Aruna Tiwari also helped me a lot with various matters related to administration.

I would like to thank all my fellow lab-mates and friends especially, Pramod Mane, Navneet Pratap Singh, Rajendra Choudhary, Chandan Gautam, Rohit Agrawal, Ram Prakash Sharma, Aaditya Chouhan for the stimulating discussions and all the fun we have had in the last few years. I specially thank **Mr. Shailendra Verma** for his help in academic and non-academic activities. I am also grateful to my friend Manish

Paliwal from NIT Nagpur for his support throughout this journey.

I gratefully acknowledge the funding received towards my Ph.D. from "Visvesvaraya Ph.D. Scheme for Electronics and IT" initiated by the Ministry of Electronics and Information Technology (MeitY).

I sincerely thanks Mrs. Richa Ahuja and Master Aarav Ahuja for their patience and compromise. I appreciate that they allowed me to have technical discussions with my supervisor during their personal time.

Last but not least, I would like to thank my family members - Mr. Anand Shastri and Mrs. Sangita Shastri (my parents), Ms. Rashmi Shastri (my sister), Mrs. Khushboo Shastri (my wife), and Ishika (my sweet little daughter) for their endless love and overwhelming support. They supported me spiritually and mentally throughout this journey and my life in general. Finally, I am thankful to all who directly or indirectly helped and supported me.

Aditya A. Shastri

To My Parents

Anand Shastri & Sangita Shastri

ABSTRACT

The two critical problems faced by the present world are depreciation in agricultural productivity and depleting human health. Specifically, due to climate change, scarcity of water and excessive heat cause decrease in the productivity of crops. Thus, the *first* part of this dissertation focuses on developing efficient variants of the standard clustering algorithm to obtain the species of crops that can be grown in less water and high heat. Furthermore, cancer has emerged as an important cause of mortality after the cardiac diseases. Hence, the *second* part focuses on developing image classification systems to accurately classify the cancer images for their early detection and prevention.

To increase the agricultural productivity, it is very important to study the genetic and phenotypic data associated with the crops (henceforth referred as plants). Genetic data is in the form of Whole Genome Sequence (WGS), which is a sequence made from a combination of four nucleotides: A (Adenine), T (Thymine), G (Guanine), and C (Cytosine). Phenotypic data are all kinds of information regarding physical characteristics of plants, such as Plant Height, 100 Seed Weight, Seed Yield Per Plant, Number of Branches Per Plant, Days to 50% Flowering, Days to Maturity, etc.

We develop a Vector Quantized Spectral Clustering (VQSC) algorithm that is a combination of Spectral Clustering (SC) and Vector Quantization (VQ) sampling for grouping genome sequences of plants. The novelty of our algorithm is in developing the crucial similarity matrix in SC as well as use of k-medoids in VQ. For genetic data of Soybean plant, we compare VQSC with commonly used techniques like Un-weighted Pair Graph Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ). Experimental results on the standard set of 31 Soybean sequences show that our VQSC outperforms both these techniques significantly in terms of cluster quality (average improvement of 21% over UPGMA and 24% over NJ) as well as time complexity (order of magnitude faster than both UPGMA and NJ).

Similarly, we develop a Probabilistically Sampled Spectral Clustering that is a combination of SC and Pivotal Sampling for grouping phenotypic data. The novelty

of our algorithm is again in constructing the crucial similarity matrix for the clustering algorithm and defining probabilities for the sampling technique. For phenotypic data of Soybean plant, we compare our algorithm with the traditional Hierarchical Clustering (HC) algorithm. Experimental results on commonly used 2400 Soybean genotypes show that we get up to 45% better quality clusters than HC in terms of Silhouette Value. Again, the complexity of our algorithm is more than a magnitude lesser than HC.

The two common cancers prevailing in the world are breast cancer and thyroid cancer. These cancers are becoming pervasive with their early detection forming a big step in saving the life of any patient. The traditional diagnostic techniques highly depend upon the personal knowledge and the experience of the doctor, where they diagnose the presence of cancerous tumor from images (X-ray image, ultrasound image, magnetic resonance image etc.). Hence, now-a-days, automated imaging techniques are commonly used for these cancer diagnosis. The most important step here is classification of the cancer images as benign or malignant. Mammography is the most effective tool for early detection of breast cancer that uses a low-dose X-ray radiation, and is commonly used. Similarly, ultrasound images (that use high frequency sound waves) of thyroid gland of a human being are mostly used for detecting thyroid cancer.

Texture of a breast and thyroid in these images plays a significant role in classifying them as benign or malignant. We propose a descriptor that is a combination of Histogram of Gradients (HOG) and Gabor filter, which exploits textural information. We term it as Histogram of Oriented Texture (HOT). We also revisit the Pass Band - Discrete Cosine Transform (PB-DCT) descriptor that captures texture information well. All features of the cancer images may not be useful. Hence, we apply a feature selection technique called Discrimination Potentiality (DP). Our resulting descriptors, DP-HOT and DP-PB-DCT, are compared with the standard descriptors. Experimental results on breast and thyroid images show that we achieve an average accuracy of 92% and 96%, respectively which is substantially more than the existing standard descriptors.

LIST OF PUBLICATIONS

Journal Papers

1. **Aditya A. Shastri**, Deepti Tamrakar and Kapil Ahuja, "Density-Wise Two Stage Mammogram Classification using Texture Exploiting Descriptors", *Expert Systems with Applications*, Elsevier, vol. 99, pp. 71–82, 2018. [**IF: 6.96**]

2. **Aditya A. Shastri**, Kapil Ahuja, Milind B. Ratnaparkhe, Aditya Shah, Aishwary Gagrani and Anant Lal, "Vector Quantized Spectral Clustering Applied to Whole Genome Sequences of Plants", *Evolutionary Bioinformatics*, SAGE, vol. 15, pp. 1–7, 2019. [**IF: 1.63**]

3. **Aditya A. Shastri**, Kapil Ahuja, Milind B. Ratnaparkhe and Yann Busnel, "Probabilistically Sampled and Spectrally Clustered Plant Species using Phenotypic Characteristics", *PeerJ*, July 2021 (Accepted). [**IF: 2.98**]

Conference Proceeding

4. Vishal Nemade, **Aditya A. Shastri**, Kapil Ahuja and Aruna Tiwari, "Scaled and Projected Spectral Clustering with Vector Quantization for Handling Big Data", *Proc. of the* 9^{th} *Symposium Series on Computational Intelligence (SSCI)*, IEEE, Bengaluru, India, pp. 2174–2179, 2018.

Contents

Abstract

List of Figures v

List of Tables vii

1 Introduction 1

1.1 Clustering Genetic Data of Plants 2

1.2 Clustering Phenotypic Data of Plants 4

1.3 Mammogram Patch Classification System 6

1.4 Thyroid Nodule Classification System 8

2 Background 11

2.1 Spectral Clustering 12

2.1.1 The Similarity Matrix 12

2.1.2 The Laplacian Matrix 12

2.1.3 The SC Algorithm 13

2.2 Vector Quantization Sampling 13

2.3 Pivotal Sampling 15

2.4 Support Vector Machine Classifier 16

3 Vector Quantized Spectral Clustering (VQSC) for Genetic Data 19

3.1 Literature Review 20

3.2 The VQSC Algorithm 21

3.3 Discussion . . . 22
3.3.1 Computational Complexity . . . 23
3.3.2 Validation Metrics . . . 23
3.4 Results . . . 24

4 Probabilistically Sampled Spectral Clustering for Phenotypic Data 33
4.1 Literature Review . . . 35
4.1.1 First Category Previous Studies . . . 35
4.1.2 Second Category Previous Studies . . . 37
4.1.3 Both Categories Previous Studies . . . 37
4.2 Our Algorithm . . . 40
4.2.1 Implementing Modified SC for Phenotypic Data . . . 40
4.2.2 Applying Pivotal Sampling to Phenotypic Data . . . 41
4.3 Results . . . 43
4.3.1 Data Description . . . 43
4.3.2 Clustering Setup . . . 44
4.3.3 Clustering and Sampling Results . . . 47
4.3.4 Sampling Estimators . . . 53

5 Classification of the Mammogram Patches 55
5.1 Literature Review . . . 56
5.2 Proposed Mammogram Patch Classification System . . . 61
5.2.1 Pre-processing and Enhancement . . . 62
5.2.2 Feature Extraction Techniques . . . 63
5.2.3 Feature Selection with Discrimination Potentiality . . . 68
5.3 Experimental Results . . . 70
5.3.1 Performance of DP-HOT . . . 72
5.3.2 Performance of DP-PB-DCT . . . 73
5.3.3 Comparison with Other Techniques . . . 74

6 Classification of the Thyroid Nodules **81**

6.1 Literature Review 82

6.2 Classification Process 83

6.3 Experimental Results 85

7 Conclusions and Future Work **89**

7.1 Clustering Algorithm Variants 89

7.2 Cancerous Image Classification System 91

A Validation of Soybean Phenotypic Data **109**

B Comparison of Pivotal Sampling with Other Samplings **111**

C Modified Spectral Clustering for Maize and Rice Phenotypic Data **113**

List of Figures

2.1 Example of an optimal hyperplane in a SVM. 17

3.1 Cluster formation for SC and VQSC with Alignment Score and $m = 11$. 28

3.2 Cluster formation for SC and VQSC with Alignment Score and $m = 12$. 28

4.1 Fifty Smallest Eigenvalues of the Type-3 Laplacian Matrix Obtained from the Euclidean Similarity Matrix (for estimating the ideal number of clusters). 45

4.2 Distribution of Genotypes (HC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size ten. 50

4.3 Distribution of Genotypes (modified SC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size ten. 50

5.1 Samples of mammogram patches from the IRMA database. Row denotes the density of patches. 60

5.2 Flow diagram of the proposed mammogram patch classification system. 61

5.3 A preprocessed and enhanced mammogram patch. 63

5.4 The HOG descriptor calculation. 65

5.5 Gabor magnitude and angle image. 66

5.6 Comparison of normal-abnormal classification accuracies obtained by varying the value of σ from one to five for each individual BIRADS class. 72

5.7 Comparison of benign-malignant classification accuracies obtained by varying the value of σ from one to five for each individual BIRADS class. 73

5.8 Performance accuracy against the number of DP-PB-DCT features. . . 76

6.1 Steps in image binarization: (a) an input ultrasound thyroid image; (b) binarized image with threshold = 10; (c) largest object detection; (d) final extracted thyroid region. 85

List of Tables

3.1 Silhouette Values for different clustering algorithms without VQ. 25

3.2 Comparison of SC with UPGMA and NJ. 26

3.3 Silhouette Values for different clustering algorithms with VQ. 27

3.4 Comparison of VQSC with VQUPGMA and VQNJ. 28

3.5 Loss in cluster quality in terms of Silhouette Values because of sampling in SC. 29

3.6 Wrongly clustered sequences by VQSC when compared with SC. 30

3.7 Comparison of VQSC with UPGMA and NJ. 31

4.1 Summary of first category previous studies. 36

4.2 Summary of second category previous studies. 38

4.3 Summary of both categories previous studies. 39

4.4 Computational complexity comparison for the given data. 44

4.5 Silhouette Values for modified SC with seven similarity measures and three Laplacian matrices for $k = 10, 20$, and 30. Silhouette Values in bold represent good clustering. 46

4.6 Loss in Silhouette Values because of Pivotal Sampling in modified SC for cluster size ten. 48

4.7 Silhouette Values for modified SC and HC with Pivotal Sampling and VQ for $N = 500$. Silhouette Values in bold represent good clustering. . 49

4.8 Silhouette Values for modified SC with Pivotal Sampling and VQ for $N = 300$. 52

4.9 Silhouette Values of modified SC with Pivotal Sampling and HC for cluster size ten. 52

4.10 HT and Hájek estimators values for Pivotal Sampling and VQ as compared to the actual population total with $N = 500$ as the sample size. . 54

5.1 Distribution of normal, benign, and malignant mammogram patches of the two different datasets for the four BIRADS classes. 59
5.2 Summary of some related works on mammogram patch classification. . 60
5.3 Comparison of various descriptors using a combination of HOG and Gabor filter 67
5.4 Feature length for the DP-HOT descriptor. 74
5.5 Feature length for the DP-PB-DCT descriptor. 75
5.6 Mammogram patch classification results as normal-abnormal for the MIAS and DDSM datasets. 78
5.7 Mammogram patch classification results as benign-malignant for the MIAS and DDSM datasets. 80

6.1 Comparison of existing and proposed classification systems. 84
6.2 Distribution of benign and malignant images according to the TIRADS classes. 86
6.3 Comparison of classification accuracies for various descriptors on the TDID dataset. 87

A.1 Phenotypic data of the Soybean genotypes used for experiments. 109
A.2 Comparison of SD, CV, mean, and range for our phenotypic data and similar previous data. 110

B.1 Comparison of Pivotal Sampling and Power Core method for three characteristics. 112

C.1 Phenotypic data of the Maize genotypes. 113
C.2 Phenotypic data of the Rice genotypes. 114
C.3 Silhouette Values of modified SC and HC for three clusters of ten Rice genotypes. 115

Chapter 1

Introduction

Depreciation in agricultural productivity and depleting human health are two critical problems currently faced by humans throughout the world. Extreme climate conditions, like drought or heat waves, cause reduced agricultural productivity [1]. This threatens the livelihoods of farmers and the food requirements of communities worldwide. One possible solution to this problem is to increase the agricultural land by cutting down the forests. However, loss of trees and other vegetations can cause climate change, including soil erosion, flooding, increased greenhouse gases, etc.

An alternative solution to this problem is to develop a better species of crops (henceforth referred as plants) having improved characteristics.[1] Thus, it is crucial to obtain the most diverse parent species that can be used for breeding. These diverse species can be obtained by studying the variation in genetic and phenotypic data associated with plants. Clustering is an important tool to analyze this variation present among different plant species. Thus, the *first* part of this dissertation focuses on developing efficient variants of the standard clustering algorithms for clustering plant species, with efficiency obtained by novel samplings. This aspect for genetic and phenotypic data for plants is briefly discussed in Sections 1.1 and 1.2 below, and further expanded in Chapters 3 and 4, respectively.

Cancer has emerged as a significant factor after the cardiac arrest that is severely affecting human health. Cancer develops when the body's cells start dividing uncon-

[1]For example, species that can grow in less water and survive high temperatures.

trollably. The five most common cancer types globally are breast cancer, oral cancer, cervical cancer, lung cancer, and stomach cancer [2]. Unlike benign cancerous tumors, malignant tumors spread to the other parts of the body and can be life-threatening. A possible solution to this problem is to classify the tumors as benign and malignant in their early stages, which can reduce the chances of death of the patient. In this dissertation, we mainly focus on the first two most common cancers, i.e. breast and oral. In the oral cancer context, we focus on thyroid cancer.

Cancerous images can be better classified by using their texture properties. Thus, the *second* part of this dissertation focuses on developing image classification systems that capture the textural features of cancerous images for their more accurate classification as benign and malignant. This aspect for breast and thyroid images is briefly discussed in Sections 1.3 and 1.4, which are further expanded in Chapters 5 and 6, respectively. The standard algorithms that are used throughout this dissertation are discussed in Chapter 2. Finally, Chapter 7 gives the concluding remarks and discusses the future works.

1.1 Clustering Genetic Data of Plants

Clustering is one of the most widely used techniques of machine learning for data analysis. People attempt to get a first impression of their data by trying to identify groups having similar behavior. Finding tight clusters, i.e. well separated and compact, is very important. Commonly used clustering algorithms include k-means, Partition Around Medoids (PAM), Clustering LARge Applications (CLARA), Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Density Based Spatial Clustering of Applications with Noise (DBSCAN), Wave-Cluster, Expectation-Maximization (EM), etc. [3]. Compared with these traditional algorithms, a promising alternative is to use spectral methods for clustering.

Clustering algorithms that use spectral properties are widely used because of their ability to generate good quality clusters (i.e. we get more tight clusters) and easy implementation (these algorithms can be solved efficiently by using standard linear

algebra methods) [4]. However, when the input data are very large, they become inefficient; computational complexity of $\mathcal{O}(n^3)$, where n is the size of the input data. Hence, considerable research has been done to reduce this complexity without affecting the accuracy of the underlying algorithm.

One such method is sampling that can reduce the input size. Samples should be selected in a manner such that they represent the whole dataset uniformly. Many techniques exist for sampling like random sampling, stratified sampling, matrix factorization, Vector Quantization (VQ), Pivotal Sampling, the strip method, the mean method, the second derivative method, etc. [5, 6]. Among these, VQ [7] is commonly used and is easy to implement because it provides the reduced data in a single scan of elements.

Clustering of Whole Genome Sequences (WGSs)[2] is useful in developing better species of plants, e.g., disease resistant and drought resistant. Here, the traditional methods for clustering, like Un-weighted Pair Graph Method with Arithmetic mean (UPGMA)[8] and Neighbor Joining (NJ)[8], which are currently used by plant biologists, do not provide good quality clusters that are needed and are also not the most efficient methods because of their high computational complexity $\mathcal{O}(n^3)$.

In this work, we present a Vector Quantized Spectral Clustering (VQSC) algorithm for grouping Single Nucleotide Polymorphism (SNP)[3] data obtained from the WGSs of plants. Although this combination of Spectral Clustering (SC) and VQ is not new [7], the novelty of our work is using the two for clustering SNP data. We test our algorithm on SNP sequences obtained from a standard plant database (Soybean) [9]. We also compare our results with currently used methods of clustering SNP data (mentioned above). Experiments show that VQSC performs better than these two popular existing techniques in terms of cluster quality (average improvement of 21% over UPGMA and 24% over NJ) as well as time complexity (order of magnitude faster than both UPGMA and NJ).

[2]A sequence made from a combination of 4 nucleotides: A (Adenine), T (Thymine), G (Guanine), and C (Cytosine).

[3]The variation in the nucleotide that occurs at a specific position across sequences.

1.2 Clustering Phenotypic Data of Plants

As mentioned earlier, variabilities present among the different plant species (also called genotypes) are useful in their breeding programs. Here, again, the selection of diverse parent genotypes is important. More diverse the parents, the higher are the chances of developing new plant varieties having excellent qualities [10]. A commonly used technique here is to study the genetic variability which, as discussed in the previous section, looks at the different genome sequences. However, this kind of analysis requires a large number of sequences, while very few are available [11, 12] because genome sequencing is computationally and monetarily expensive [13].

Variabilities in plant genotypes can also be studied using their phenotypic characteristics (physical characteristics). This kind of analysis can be relatively easily done because a sufficiently large amount of data is available from different geographical areas. In the phenotypic context, a few characteristics that play an important role are Days to 50% Flowering, Days to Maturity, Plant Height, 100 Seed Weight, Seed Yield Per Plant, Number of Branches Per Plant, etc.

As discussed for the genetic data earlier, cluster analysis is an important tool to describe and summarize the variation present between different plant genotypes using the phenotypic data as well [10]. This data for the genotypes of different plants (e.g., Soybean, Wheat, Rice, Maize, etc.) usually have enough variation for better clustering. However, if this data is obtained for the genotypes of the same plant, then clustering becomes challenging due to less variation in the data, which forms our focus.

Hierarchical Clustering (HC) is a traditional and standard method that is currently being used by plant biologists for grouping of phenotypic data [10, 14, 15]. However, this method has a few disadvantages. First, it does not provide better quality clusters when grouping similar genotypes [16]. Second, HC is based on building a hierarchical cluster tree (also called dendrogram), which becomes cumbersome and impractical to visualize when the data is too large.

To overcome these two disadvantages of HC, we propose the use of the SC algorithm. SC is mathematically sound and is known to give the best quality cluster

among the existing clustering algorithms [17]. As discussed in previous section, we get substantial improvements in cluster quality by using SC for genetic data. Furthermore, unlike HC, SC does not generate the intermediate hierarchical cluster tree. To the best of our knowledge, this algorithm has not been applied to phenotypic data in any of the previous works (see Section 4.1).

HC, as well as SC, both are computationally expensive. They require substantial computational time when clustering large amounts of data [17, 16]. Hence, we again use sampling to reduce this complexity. Probability-based sampling techniques have recently gained a lot of attention because of their high accuracy at reduced cost [5]. Among these, Pivotal Sampling is most commonly used [18], and hence, we apply it to phenotypic data. Like for SC, using Pivotal Sampling for phenotypic data is also new. Recently, VQ has given promising results for genetic data (discussed above). Hence, here we adapt VQ for phenotypic data as well. This also serves as a good standard against which we compare Pivotal Sampling.

To summarize, we develop a modified SC with Pivotal Sampling algorithm that is especially adapted for phenotypic data. The novelty of our work is in constructing the crucial similarity matrix for the clustering algorithm and defining the probabilities for the sampling technique. Although our algorithm can be applied to any plant genotypes, we test it on around 2400 Soybean genotypes obtained from Indian Institute of Soybean Research, Indore, India [19]. In the experiments, we perform four sets of comparisons. First, we show that use of Pivotal Sampling does not deteriorate the cluster quality. Second, our algorithm outperforms all the proposed competitive clustering algorithms with sampling in terms of cluster quality (i.e. modified SC with VQ, HC with Pivotal Sampling, and HC with VQ). The computational complexities of all these algorithms are similar because of the involved sampling. Third, our modified SC with Pivotal Sampling doubly outperforms HC, which as earlier, is a standard in the plant studies domain. In terms of Silhouette Value, we get up to 45% better quality clusters. In terms of complexity, our algorithm is more than a magnitude cheaper than HC. Fourth and finally, we demonstrate the superiority of our algorithm by comparing it with two previous works that are closest to ours.

1.3 Mammogram Patch Classification System

Breast cancer has become the most common killer disease in the female population. Collectively India, China and US have almost one-third burden of global breast cancer [20]. The abnormalities like the existence of a breast mass, change in shape, the dimension of the breast, differences in the color of the breast skin, breast aches, etc., are the symptoms of breast cancer. Cancer diagnosis is performed based upon non-molecular criteria like the tissue type, pathological properties and the clinical location. Cancer begins with the uncontrolled division of one cell and results in the form of a tumor.

There are several imaging techniques for examination of the breast, such as magnetic resonance imaging, ultrasound imaging, X-ray imaging, etc. Mammography is the most effective tool for early detection of breast cancer that uses a low-dose X-ray radiation. It can reveal pronounce evidence of abnormalities, such as masses and calcification, as well as subtle signs, such as bilateral asymmetry and architectural distortion. The diagnosis of breast cancer by classifying it as benign and malignant in the early stage can reduce chances of the death of the patient.

Mammographic Computer Aided Diagnosis (CAD) systems enable evaluation of abnormalities (e.g., micro-calcification, masses, and distortions) in mammography images. CAD systems are necessary to aid facilities in carrying out a more accurate diagnosis. CAD systems are designed with either fully automatic or semi-automatic tools to assist radiologists for detection and classification of mammography abnormalities [21]. In semi-automated CAD systems, enhancement techniques are first applied on a mammogram patch, radiologists then select a Region of Interest (ROI) or a patch, and finally, the patch is classified by the system.

Mammogram patch classification is often done in one stage. However, classifying a mammogram patch in multiple stages is also beneficial. Two-stage classification of mammogram patches helps in reducing the possibility of a false positive classification. In the first stage, mammogram patches are classified as normal or abnormal, then in the second stage, abnormal patches are further classified into benign or malignant.

This work proposes two-stage mammogram patch classification. The system is trained with normal, benign and malignant mammogram patches separately.

Generally, Computer Aided Diagnosis (CAD) systems consist of basic modules as follows: mammogram patch pre-processing, breast segmentation, enhancement, feature extraction and classification [22]. Pre-processing step helps in removal of irrelevant regions present in a mammogram patch such as pectoral muscles and digit information. Breast region is segmented using a threshold. Enhancement techniques such as adaptive histogram equalization, non-linear filtering are applied on the breast region to improve visualization of tissues or a tumor in a mammogram patch [23, 24, 25]. In most works, shape features of a mammogram patch have only been considered. The shape of a mammogram patch plays an important role for benign and malignant classification. While benign masses have round or oval shapes with clear margins, malignant masses with spicule have jagged edges [26]. Appropriate features of mammogram patches help in accurate classification.

Mammogram patches can be better classified by using their texture properties. This work proposes a descriptor that captures the textural features of a mammogram patch, i.e. Histogram of Oriented Texture (HOT), which is a variation of Histogram of Gradients (HOG) and Gabor filter combination. We also apply the existing Pass Band - Discrete Cosine Transform based descriptor (PB-DCT) here because of its advantage in helping filter textural features. These descriptors have not been used yet for mammogram patch classification. We use Discrimination Potentiality (DP) to select appropriate features of mammogram patches in these two descriptors, resulting in two new descriptors (DP-HOT and DP-PB-DCT). The proposed descriptors are compared with the six standard descriptors for mammogram patch classification; Zernike moments [27], MLPQ [28], GRsca [29], Wavelet Gray Level Co-occurrence Matrix (WGLCM) [30], Local Configure Pattern (LCP) and HOG [31]. SVM is the most suitable classifier for two-class classification and is widely used in this field. Hence, we use this.

Breasts with high density have a higher chance of cancer. However, high dense tissues and masses appear as mostly white in a gray scale of a mammogram patch.

Hence, it is very difficult to detect a tumor in high dense tissues. Especially, the difference between benign and malignant tumors is hard to determine [21, 32, 33, 34, 35]. Generally, breasts are classified based upon density in three different ways by the Breast Imaging Reporting And Database Systems (BIRADS); two classes (fatty and dense), three classes (fatty, glandular, and dense) or four classes (mostly fatty, scattered density, consistent density and extremely dense) [32, 33]. Most researchers in this area have not considered the density of a breast for mammogram patch classification. Hence, in this work, we test our two proposed descriptors for each BIRADS class separately and combined.

CAD systems are usually tested on the MIAS and DDSM mammogram patch datasets of the IRMA database [36]. The MIAS dataset consists of a small set of images, while DDSM includes few thousand images. Several descriptors and methodologies have been proposed for mammogram patch classification, but their performances have been investigated only for a small set of images. Moreover, these systems have not achieved desired accuracy [32, 35]. The performance of our system is tested on all mammogram patches of the MIAS and DDSM datasets. The experimental results show the effectiveness of our approach as we achieve near to 92% accuracy.

1.4 Thyroid Nodule Classification System

Thyroid nodule (or a lump) that develops in the thyroid gland of a human being, is a disease in which cells grow abnormally and are likely to spread to the other parts of the body [37]. Presence of this nodule may or may not be an indication of thyroid cancer. When a thyroid nodule is found, scanning/ imaging of the thyroid region is done to check if this nodule is a benign or a malignant nodule. Favorably, most of the detected thyroid nodules are benign. However, the presence of a nodule (whether benign or malignant) causes various health problems in patients like difficulty in breathing and swallowing [38]. Moreover, malignant thyroid nodules can produce an additional hormone called thyroxine, which causes some critical problems with patient's health and may result in his/ her death [38]. Hence, classifying these nodules

at an early stage can reduce chances of the death of the patient.

Abnormalities like hoarseness, swollen glands in the neck, difficulty in swallowing, difficulty in breathing, pain in the throat or neck, a lump in the front of the neck (near the Adam's apple), etc. are some of the symptoms of thyroid cancer [37]. There exist several imaging techniques for examination of the thyroid, such as computed tomography scanning, ultrasound imaging, X-ray imaging, etc. [37]. Ultrasound imaging is the most effective tool for an early detection of thyroid cancer that uses high-frequency sound waves to create a picture of the internal organs [39, 37, 38, 40].

The traditional diagnostic technique, where doctors diagnose the presence of cancerous tumor from the ultrasound images, may give false results as this diagnosis heavily relies upon the personal knowledge and the experience of the doctor. That is, determining whether a thyroid nodule is benign or malignant is a hard task for doctors as well because it is based only upon symptoms and/ or experience. Hence, now-a-days, researchers are focusing on developing Artificial Intelligence (AI) based imaging techniques for this purpose [39, 38]. Development of an image-based Computer-Aided Diagnosis (CAD) system in medical research serves as an additional expert that assists doctors in accurate diagnosis.

Similar to the mammographic CAD system, CAD system for thyroid nodule classification also consists of following basic modules: pre-processing thyroid ultrasound images, image enhancement, feature extraction and classification. Here, pre-processing step helps in removal of background and artifacts (additional text or indicator made by the capturing system). Enhancement techniques are applied on the thyroid images to improve visualization of tissues or a tumor. Finally, a feature extraction technique is employed to obtain the features from images and a classifier to classify them.

Similar to the mammogram patches, thyroid images can also be better classified by using their texture properties [41]. Hence, for classifying thyroid nodules, we again propose use of the two descriptors that capture the textural features, i.e. HOT and PB-DCT. These descriptors have not been used yet for this type of classification. As mentioned earlier, we use DP to select the appropriate features.

A few characteristics of thyroid nodules from the ultrasound image are used as

suggestive features for malignancy. These include micro-calcifications, absence of a halo, solidity, intra-nodular flow, hypo-echogenicity and taller-than-wide shape [42]. Accordingly, thyroid nodules are classified into following categories by the Thyroid Imaging Reporting And Data System (TIRADS) [43]; not suspicious, probably benign, one suspicious features, two suspicious features, three or more suspicious features and probable malignancy. These categories are represented by the TIRADS scores of 2, 3, 4a, 4b, 4c and 5, respectively. Based upon this, we consider the ultrasound images with TIRADS scores of 2 or 3 as the benign cases, while the ultrasound images with TIRADS scores of 4a, 4b, 4c and 5 as the malignant cases. Support Vector Machine (SVM) is the most suitable and widely used classifier for the two-class classification problem. Hence, we use this.

CAD systems are usually tested on the Thyroid Digital Image Dataset (TDID), an open access database of thyroid ultrasound images created by Universidad Nacional de Colombia [44]. Several methodologies have been proposed for thyroid nodules classification, but these systems have not achieved the desired accuracy [39]. That is, image augmentation [45], VGG-16 [46], GoogLeNet [47], Circular Mask [39] and Convolutional Neural Network (CNN) [39]. The performance of our classification system is tested on all thyroid ultrasound images of TDID. The experimental results show the effectiveness of our approach; we achieve near to 96% accuracy.

Chapter 2

Background

Clustering and classification are the machine learning techniques used to categorize the input instances into one or more groups/ classes based upon their features [48]. There is a fundamental difference between these two techniques. In clustering, instances are grouped based upon their similarities without having any prior information about the resulting groups. Thus, this technique comes under the unsupervised learning category. On the other hand, in classification, we classify the instances based upon their corresponding predefined class labels, and hence, this comes under the supervised learning category.

As mentioned in Chapter 1, we mainly focus on Spectral Clustering (SC) in this dissertation. Hence, the standard algorithm for SC is given in Section 2.1 of this chapter. As also mentioned earlier, to reduce the complexity of our SC variants, we use two sampling techniques: Vector Quantization (VQ) and Pivotal Sampling. These techniques are explained in Sections 2.2 and 2.3, respectively. Finally, for classification, as stated earlier as well, we use Support Vector Machine (SVM), which is the most commonly used classifier for the two-class classification (as needed here). This is explained in Section 2.4.

2.1 Spectral Clustering

SC algorithms are widely used because they generate better quality clusters (we get more tight and well separated clusters) and easy implementation (these algorithms can be solved efficiently by using standard linear algebra methods) [17, 4, 49]. Next, we provide a brief introduction to the mathematical aspects used by SC: the similarity matrix in Section 2.1.1 and the Laplacian matrix in Section 2.1.2. Finally, we present the standard algorithm for SC in Section 2.1.3.

2.1.1 The Similarity Matrix

The first step in the SC algorithm is the construction of a matrix called the similarity matrix. Building this matrix is the most important aspect of this algorithm; better its quality, better the quality of clusters. This matrix captures the local neighborhood relationships between the instances via similarity graphs and is usually built in three ways [17]. The first such graph is a ϵ-neighborhood graph, where all the vertices whose pairwise distances are smaller than ϵ are connected. The second is a k-nearest neighborhood graph, where the goal is to connect vertex v_i with vertex v_j if v_j is among the k-nearest neighbors of v_i. The third and the final is the fully connected graph, where each vertex is connected with all the other vertices. Similarities are obtained between the connected vertices only. Thus, similarity matrices obtained by the first two graphs are usually sparse, while the fully connected graph yields a dense matrix.

We can use any of the graphs mentioned above to generate the similarity matrix for SC. Since no theoretical analysis is available for helping chose a particular type of the similarity graph [17], throughout this dissertation, we use the fully connected graph.

2.1.2 The Laplacian Matrix

The next important aspect of SC is the Laplacian matrix, which is constructed from the similarity matrix. This matrix is either non-normalized or normalized. The

non-normalized Laplacian matrix is defined as

$$L = D - S,$$

where S is the similarity matrix and D is a diagonal matrix whose elements are obtained by adding together the elements of all the columns for every row of S.

Normalized Laplacian matrix is again of two types: the symmetric Laplacian (L_{sym}) and the random walk Laplacian (L_{rw}). Both these matrices are closely related to each other and are defined as

$$L_{sym} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}SD^{-1/2} \text{ and}$$

$$L_{rw} = D^{-1}L = I - D^{-1}S,$$

where I is the identity matrix. In literature, it is suggested to use the normalized Laplacian matrix instead of the non-normalized one, and specifically the random walk Laplacian matrix [17]. Hence, in this dissertation, we use this Laplacian matrix.

2.1.3 The SC Algorithm

Corresponding to the three Laplacian matrices discussed above, three variants of the SC algorithms have been successfully used in literature. That is, there exist two variants of the normalized SC and a non-normalized SC. Here, we present the normalized SC algorithm proposed by Shi and Malik (2000) [49] that is most commonly used (see Algorithm 1). This algorithm uses the random walk Laplacian matrix as discussed above. We refer the readers to [17] for non-normalized SC and [4] for normalized SC using L_{sym}.

2.2 Vector Quantization Sampling

VQ is a data compression technique that encodes/ maps each input data (also called input vector) to its closest matching representative vector [7]. The most important component of VQ is a codebook that contains the set of representative vectors [51].

Algorithm 1 Normalized SC by Shi and Malik (2000) [49]

Input: Similarity function defined in [50] and number k, which denotes the number of clusters. ▷ Here, we consider that the data consists of n instances $x_1, ..., x_n$. We obtain the similarities $s_{ij} = s(x_i, x_j)$ for $i, j = 1, ..., n$ by using a similarity function defined in [50]. The corresponding similarity matrix is denoted by $S = (s_{ij})_{i,j=1,...,n}$.

1: Construct a similarity graph by one of the ways described in Section 2.1.1. Let S be its similarity matrix.

2: Compute the non-normalized Laplacian L discussed in Section 2.1.2.

3: Compute the first k generalized eigenvectors $u_1, ..., u_k$ of the generalized eigenproblem $Lu = \lambda Du$.

4: Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, ..., u_k$ as columns.

5: For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i^{th} row of U.

6: Cluster the points of y_i with the k-means algorithm into clusters $A_1, ..., A_k$.

Output: Clusters $C_1, ..., C_k$ with $C_i = \{j | y_j \in A_i\}$.

This codebook is designed in such a manner that the difference between the original and the representative set is minimized. The most common algorithm for designing this codebook is the Linde-Buzo-Gray algorithm or the Generalized Lloyd algorithm [52]. This algorithm uses the traditional k-means clustering approach to obtain the representative vectors.

Given a set of data points $x_1, x_2, ..., x_n$, where each $x_i \in \mathbb{R}^d$ (i.e. d-dimensional vector), k-means randomly selects $m_1^{(1)}, m_2^{(1)}, ..., m_k^{(1)}$ as the initial means, where superscript denotes the iteration number and subscript the cluster index. The distances between all the n data points and these k means are calculated and each data point is assigned to the nearest mean.[1] Then, updated means $m_1^{(2)}, m_2^{(2)}, ..., m_k^{(2)}$ are obtained for each cluster using the data points assigned to that particular cluster. This algorithm converges when the change in the means is less than a certain tolerance. Finally, $m_1^{(t)}, m_2^{(t)}, ..., m_k^{(t)}$ are selected as the set of representative vectors, where t denotes the number of iterations required for convergence.

[1] These distances can be calculated by using any of the popular distance measures, e.g., Euclidean, Square Euclidean, Cosine Distances, etc. [50].

2.3 Pivotal Sampling

This is a well-developed sampling theory that handles complex data with unequal probabilities. The method is attractive because it can be easily implemented by a sequential procedure, i.e. by a single scan of the data [53]. Thus, the complexity of this method is $\mathcal{O}(n)$, where n is the population size. This can also be applied to streaming data where we do not have the list of all the units at the beginning of the sampling process. It is important to emphasize that this method is independent of the density of the data.

Consider a finite population U of size n with its each unit identified by a label $i = 1, 2, ..., n$. A sample S is a subset of U with its size, either being random ($N(S)$) or fixed (N). Like any sampling algorithm with unequal probabilities, this technique requires that the inclusion probabilities of all the units in the population, denoted by π_i with $i = 1, 2, ..., n$, may be computed before a unit is first considered for a contest. This forms an important aspect of this unequal probability sampling technique. To select a sample of size N, where $N \ll n$, we obtain these probabilities as [53]

$$\pi_i = N \frac{\varkappa_i}{\sum_{i \in U} \varkappa_i},$$

where $\varkappa_i$ can be a property associated with the data. Obtaining π_i in such a way also ensures that $\sum_{i=1}^{n} \pi_i = N$, i.e. we get exactly N selection steps (discussed next), and in turn, exactly N samples.

As above, this method is based on a principle of contests between units [5]. At each step of the method, two units compete to get selected (or rejected). Consider unit i with probability π_i and unit j with probability π_j, then we have the two cases as below.

1. **Selection step ($\pi_i + \pi_j \geq 1$):** Here, one of the units is selected, while the other one gets the residual probability $\pi_i + \pi_j - 1$ and competes with another unit at the next step. More precisely, if (π_i, π_j) denotes the selection probabilities of the two units, then

$$(\pi_i, \pi_j) = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j} \end{cases}$$

2. **Rejection step ($\pi_i + \pi_j < 1$):** Here, one of the units is definitely rejected (i.e. not selected in the sample), while the other one gets the sum of the inclusion probabilities of both the units and competes with another unit at the next step. More precisely,

$$(\pi_i, \pi_j) = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \frac{\pi_j}{\pi_i+\pi_j} \\ (\pi_i + \pi_j, 0) & \text{with probability } \frac{\pi_i}{\pi_i+\pi_j} \end{cases}$$

This step is repeated for all the units present in the population until we get the sample of size $N(S)$ or N. The worst-case occurs when we obtain the last sample (i.e. N^{th} sample) in the last iteration.

2.4 Support Vector Machine Classifier

SVM is a supervised learning technique that is used to solve both pattern classification and nonlinear-regression problems [54]. Mostly, it is used to perform two-class classification. Consider a set of training instances, where each instance belongs to one of the two classes. SVM uses the training instances to construct an optimal hyperplane in such a manner that the margin of separation (or gap) between the two classes is maximized.

While testing, new instances are classified based on which side of the hyperplane they fall. Figure 2.1 gives an example of an optimal hyperplane that separates the instances of two classes (represented as circles and squares). Here, the instances colored in blue are the support vectors and d is the distance between the optimal hyperplane and the nearest training instance of a class (or support vector). Besides linear classification, SVM performs non-linear classification as well where a kernel trick is used. Here, we only need to perform linear classification, and hence, we use a linear SVM.

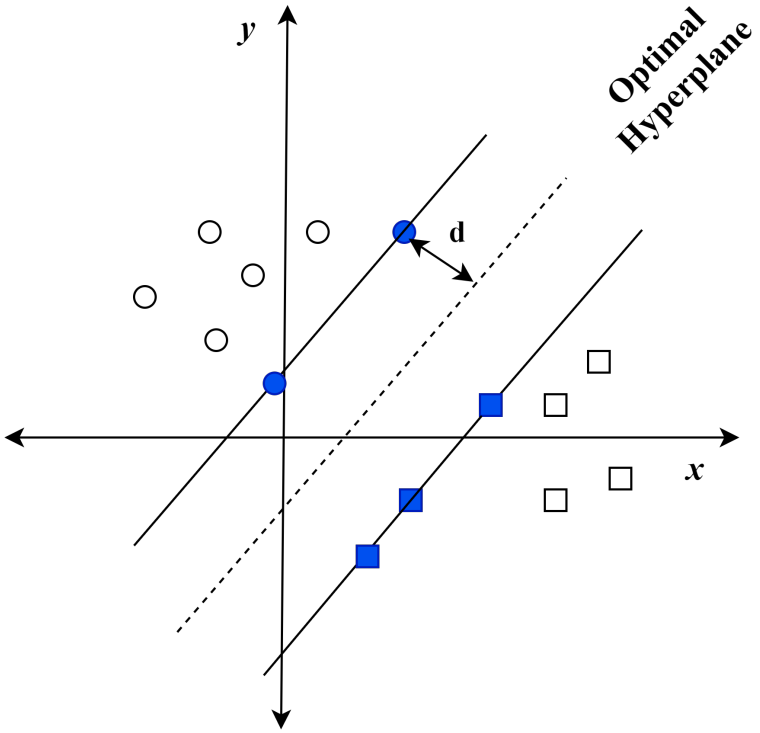


Figure 2.1: Example of an optimal hyperplane in a SVM.

Chapter 3

Vector Quantized Spectral Clustering (VQSC) for Genetic Data

Every specie of a plant is genetically represented as combinations of four nucleotides: A (Adenine), T (Thymine), G (Guanine), and C (Cytosine), which are called the Whole Genome Sequences (WGSs). Every WGS is very long consisting of several billions of such nucleotides. Hence, typically the size of each WGS runs into tens of GBs. Single Nucleotide Polymorphisms (SNPs) are the variations present in different WGSs at the nucleotide level. Instead of working with WGSs, scientist often work with SNPs because these variations affect how plants develop diseases and respond to pathogens, chemicals, drugs etc. [55].

Out of all the available WGSs (or SNPs) for different species of a plant, only a few are known to have a particular trait, e.g., disease resistant, drought resistant etc. Clustering of these WGSs (or SNPs) is very useful in determining the similar sequences, which in turn can be used to develop plant species with a combination of useful traits. Hence, we need better quality clustering of these WGSs (or SNPs), i.e. the clusters obtained should be compact and well separated from each other. This should be done efficiently as well.

In this chapter, we present a Vector Quantized Spectral Clustering (VQSC) algo-

rithm that is a combination of Spectral Clustering (SC) and Vector Quantization (VQ) sampling for grouping genome sequences of plants. The inspiration here is to use SC that gives good quality clusters and VQ to make the algorithm computationally cheap (the complexity of SC is cubic in terms of the input size). Although the combination of SC and VQ is not new, the novelty of our work is in developing the crucial similarity matrix in SC as well as use of k-medoids in VQ, both adapted for the plant genetic data. For Soybean genome sequences, we compare our approach with commonly used techniques like Un-weighted Pair Graph Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ). Experimental results show that our VQSC outperforms both these techniques significantly in terms of cluster quality (average improvement of 21% over UPGMA and 24% over NJ) as well as time complexity (order of magnitude faster than both UPGMA and NJ).

3.1 Literature Review

Here, we present literature regarding usage of SC and VQ in the field of plant genome, and the novelty of our approach. SC can be performed in two ways: recursive and non-recursive. Bouaziz et al. [56] in 2012 used this method in a recursive way for genetic studies. However, we use a common non-recursive way [17, 4], because it is simpler and cheaper. It also gives tight and compact clusters.

The construction of the similarity matrix is the most important part of the SC algorithm. This can be done either by using basic techniques like pairwise distance [57], Jukes Cantor [58], Alignment Score [59], cosine similarity and others [50], or by using advanced techniques [60] like identity-by-state, allele sharing distance, SNP edit distance, covariance, normalized covariance, and coancestry.

Li et al. [61] in 2010 used SC for clustering gene sequences (which are a subset of WGSs) where they constructed the similarity matrix by cosine similarity. We use the earlier mentioned basic techniques besides cosine similarity because they capture the similarity between the SNP sequences in a better way. We do not use advanced techniques because they are more involved (and also not needed since basic work well).

Zhang et al. [62] in 2011 used VQ to reduce the number of genome sequences of influenza A virus for better visualization of phylogenetic trees, which is an essential step in earlier mentioned clustering algorithms of UPGMA and NJ. They used the neural gas method as the basis of their sampling.

We use VQ as well, but in a different sense. We use k-medoids as the basis of our sampling instead of the neural gas method. This is because it is easy to find the medoids of the kind of data we have.

3.2 The VQSC Algorithm

As mentioned earlier, construction of the similarity matrix is significant in the SC algorithm because better the quality of this matrix, better is the quality of clusters generated by this algorithm. Hence, in this work, for constructing the similarity matrix, we compare every character in one SNP sequence with every character in other SNP sequences. This represents how much one sequence is different from another sequence. The dissimilarity $D(i, j)$ between any two SNP sequences X_i and X_j is defined as the number of positions at which X_i and X_j differ. The similarity value is calculated as

$$S(i, j) = l(seq) - D(i, j),$$

where, $l(seq)$ is the length of the SNP sequence. This value is normalized and used as the similarity value for (i, j) index. We also use other similarity measures like pair-wise distance, Jukes Cantor, and Alignment Score to construct the similarity matrix. Results show that the quality of clusters is sensitive to the quality of the similarity matrix used.

As mentioned earlier, we use VQ to compress the original data into a small set of representative data entities. The goal now is to minimize the difference between the original and this representative set. Although the standard VQ algorithm uses k-means, we achieve this minimized difference by using the k-medoids algorithm. This is because, as discussed earlier, data here are in the form of sequences of strings of A, T, G, and C characters and mean of these data does not exist. On the other hand,

k-medoids provide us with representative sequences from the set of given sequences itself. We briefly summarize VQSC in Algorithm 2 below.

Algorithm 2 The VQSC Algorithm

Input: n SNP sequences $\{x_i\}$ for $i = 1, ..., n$; k number of sample sequences to be selected; and m number of clusters to be formed.

1: Perform k-medoids as follows:

(a) Compute medoids $y_1, ..., y_k$ as the k sample sequences.

(b) Build a correspondence table to associate each x_i with the nearest medoid y_j.

2: Run the SC algorithm on $y_1, ..., y_k$ to obtain cluster indexes C_l; $l = 1, ..., m$ for each of y_j.

3: Recover the cluster membership for each x_i by looking up the correspondence table.

Output: clustered SNP sequences.

3.3 Discussion

We use SNP data of 31 Soybean sequences, which are taken from the database as follows: http://chibba.pgml.uga.edu/snphylo/ [9]. These data contain 62,89,747 SNPs. As this is a raw data, we use SNPhylo software [9] to remove low-quality data. Specifically, false SNPs are removed and we get 31 SNP sequences each of length 4847.[1] Please refer to Figure 1 of Lee et al. [9], which shows the flowchart of SNPhylo pipeline, which is a commonly used standard procedure. Finally, these sequences are used to obtain the similarities among each other leading to the construction of the similarity matrix, which is an input to our VQSC algorithm.

Next, we first discuss the computational complexity of our and other standard algorithms (for SNP clustering). Then, we describe the criteria used to check the

[1]This software also constructs a phylogenetic tree as used by other standard genome clustering algorithms.

goodness of generated clusters, termed as validation metrics.

3.3.1 Computational Complexity

As mentioned in Chapter 1, complexities of the standard SC, UPGMA, and NJ algorithms are all $\mathcal{O}(n^3)$, where n is the size of the input data. This makes these algorithms computationally less efficient. However, the use of VQ sampling with SC reduces the complexity of VQSC to $\mathcal{O}(k^3 + n^2kt)$, where k is the number of representative samples chosen via k-medoids in VQ, and t is the number of iterations taken by VQ. Here, the first term (k^3) comes from SC, and the second term (n^2kt) comes from VQ. Application of VQ to UPGMA and NJ also leads to a comparable reduction in their complexities.

3.3.2 Validation Metrics

There are various metrics available for the validation of clustering algorithms. These include Cluster Accuracy (CA), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Compactness (CP), Separation (SP), Davis-Bouldin Index (DB), and Silhouette Value [3, 63]. For using the first three metrics, we should have a prior knowledge of the cluster labels. However, here we do not have this information. Hence, we cannot use these validation metrics. Rest of the techniques do not have this requirement, and hence, can be used for validation here. We use Silhouette Value because of its popularity [63].

Silhouette Value is a measure of how similar an object is to its own cluster (intra-cluster similarity) compared with other clusters (inter-cluster similarity). For any cluster C_l ($l = 1, ..., k$; say $l = 1$), let $a(i)$ be the average distance between the i^{th} data point and all other points in the cluster C_1, and let $b(i)$ be the average distance between this i^{th} data point in the cluster C_1 and all other points in clusters $C_2, ..., C_k$. Silhouette Value for the i^{th} data point is defined as [63]

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3.1}$$

where, $a(i)$ and $b(i)$ signify the intra-cluster and the inter-cluster similarities, respectively. Silhouette Value lies between -1 and 1,[2] and average over all the data points is computed. A positive value (tending towards 1) indicates good clustering (compact and well-separated clusters), while a negative value (tending towards -1) indicates poor clustering.

3.4 Results

We first present the results of SC, UPGMA, and NJ without VQ. These data are given in Table 3.1. Column 1 gives the number of clusters chosen. As 2 to $\frac{n}{2}$ clusters, where n is the number of input data points, are commonly used in literature, we follow this. Hence, we provide results from 2 to 16 clusters (for us $n = 31$, and hence, $\frac{n}{2} = 15.5 \approx 16$). Columns 2 to 5 refer to the Silhouette Values of the SC algorithm with four different similarity measures discussed earlier. Columns 6 and 7 give the Silhouette Values for UPGMA and NJ. As evident (highlighted in bold), SC with Alignment Score gives the best results for all the clusters. We also obtain the ideal number of clusters using eigenvalue gap heuristic [17, 74]. It comes out to be m=2.

The percentage improvement in SC (using Alignment Score as the similarity measure) in comparison with UPGMA and NJ is given in Table 3.2. We can observe from this table that the average improvement in SC over UPGMA is around 34% and over NJ is around 37%, which is considered to be a substantial improvement.

Next, we discuss the results for the same three clustering algorithms with VQ. Results for these experiments are given in Table 3.3 (structure of which is similar to that of Table 3.1). From this table, we see a similar pattern, i.e. our VQSC algorithm with Alignment Score is the best (highlighted in bold) as compared to Vector Quantized UPGMA (VQUPGMA) and Vector Quantized NJ (VQNJ).

We also calculate percentage improvement in VQSC over VQUPGMA and VQNJ. The data for this is given in Table 3.4. Again, we observe substantial improvement by using VQSC. The average percentage improvement in VQSC over VQUPGMA is

[2]This is because the denominator of (3.1) is always greater than its numerator.

Table 3.1: Silhouette Values for different clustering algorithms without VQ.

# of Clusters	**SC**				**UPGMA**	**NJ**
	Similarity $S(i,j)$	**Pairwise Distance**	**Jukes Cantor**	**Alignment Score**		
2	0.2012	0.2012	0.2590	**0.3169**	0.1831	0.2206
3	0.1987	0.1722	0.2440	**0.2845**	0.2002	0.2258
4	0.2053	0.2037	0.2621	**0.3241**	0.2546	0.2192
5	0.2488	0.2421	0.3017	**0.3528**	0.2791	0.2488
6	0.2771	0.2771	0.3214	**0.3886**	0.2389	0.2771
7	0.2990	0.3231	0.3414	**0.3882**	0.2612	0.2736
8	0.3451	0.3451	0.3811	**0.4007**	0.2906	0.2874
9	0.3490	0.3140	0.3785	**0.4130**	0.3112	0.3031
10	0.3522	0.3507	0.3771	**0.4464**	0.3430	0.2966
11	0.3687	0.3681	0.4045	**0.4589**	0.3831	0.3476
12	0.3799	0.4046	0.4258	**0.5031**	0.4089	0.3569
13	0.4329	0.3948	0.4611	**0.5375**	0.4153	0.3829
14	0.4470	0.4527	0.4646	**0.5415**	0.4610	0.4403
15	0.4481	0.4590	0.5093	**0.5701**	0.4881	0.4366
16	0.5014	0.5134	0.5301	**0.5917**	0.5139	0.4665

around 28% and over VQNJ it is around 347%.

Next, we calculate the loss in the cluster quality incurred in terms of Silhouette Value because of sampling in the proposed SC algorithm (with Alignment Score as the similarity measure). For this, we compare the relevant SC and VQSC data from Tables 3.1 and 3.3, respectively. This loss for the different number of clusters chosen is listed in Table 3.5. We can observe from these data that the average of the loss in terms of Silhouette Values comes around 11%, which is considered acceptable because we are still better than the existing best algorithms (UPGMA and NJ; please see Table 7 and the accompanying discussion below).

Table 3.2: Comparison of SC with UPGMA and NJ.

# of Clusters	**% improvement in SC**	
	Over UPGMA	**Over NJ**
2	73.07	43.65
3	42.11	26.00
4	27.30	47.86
5	26.41	41.80
6	62.66	40.24
7	48.62	41.89
8	37.89	39.42
9	32.71	36.26
10	30.15	50.51
11	19.79	32.02
12	23.04	40.96
13	29.42	40.38
14	17.46	22.98
15	16.80	30.58
16	15.14	26.84
Average	**33.50**	**37.43**

We further validate the quality of these clusters using tools used by biologists at Indian Institute of Soybean Research, Indore, India. Here, we compare cluster formation for SC and VQSC for the different number of clusters. As above, we use the data corresponding to Alignment Score as the similarity measure because that gives the best results.

We do this comparison in two ways. For two cases ($m = 11$ and 12), we diagrammatically identify the sequences that are wrongly clustered by VQSC as compared with SC (in Figures 3.1 and 3.2). For all other values of m, we give the number of sequences wrongly clustered by VQSC as compared with SC (in Table 3.6). This

Table 3.3: Silhouette Values for different clustering algorithms with VQ.

# of Clusters	**VQSC**				**VQ-UPGMA**	**VQNJ**
	Similarity $S(i,j)$	**Pairwise Distance**	**Jukes Cantor**	**Alignment Score**		
2	0.2012	0.2012	0.2590	**0.3169**	0.1835	0.0128
3	0.2002	0.2002	0.2474	**0.2876**	0.2002	0.0427
4	0.2159	0.2181	0.2610	**0.3052**	0.2192	0.0752
5	0.2211	0.2488	0.2887	**0.3232**	0.2488	0.0827
6	0.2639	0.2528	0.2922	**0.3046**	0.2532	0.0476
7	0.2446	0.2184	0.2867	**0.3259**	0.2604	0.0821
8	0.2727	0.2718	0.3189	**0.2935**	0.2752	0.1195
9	0.2861	0.3209	0.2890	**0.4004**	0.2886	0.1506
10	0.3361	0.2429	0.3561	**0.3726**	0.3264	0.1523
11	0.3035	0.2877	0.3672	**0.4594**	0.3456	0.2273
12	0.3299	0.3783	0.4078	**0.4743**	0.3650	0.2513
13	0.4268	0.4184	0.3811	**0.4843**	0.4216	0.3002
14	0.4128	0.4251	0.4450	**0.4966**	0.4111	0.3465
15	0.4560	0.4592	0.4796	**0.5334**	0.4592	0.3552
16	0.4552	0.4434	0.4587	**0.5004**	0.4434	0.4434

two-way strategy comprehensively depicts the goodness of VQSC.

In Figures 3.1 and 3.2, the x-axis lists the 31 sequences and the y-axis refers to the clustering algorithms used. The Silhouette Values from Tables 3.1 and 3.3 are given on the right. The different colors denote the different clusters, and the colored boxes signify which cluster each sequence belongs to. From Figure 3.1, we observe that our VQSC algorithm does not cluster sequences W08, W11, and C01 (i.e. only 3 out of 31) in their respective clusters when compared with SC. Similar behavior can be observed from Figure 3.2. Sequences W05, C01, and C19 (again only 3 out of 31) are not correctly clustered by VQSC when compared with SC.

Table 3.4: Comparison of VQSC with VQUPGMA and VQNJ.

# of Clusters	% improvement in VQSC	
	Over VQUPGMA	**Over VQNJ**
2	72.70	2375.78
3	43.66	573.54
4	39.23	305.85
5	29.90	290.81
6	20.30	539.92
7	25.15	296.95
8	6.65	145.61
9	38.74	165.87
10	14.15	144.65
11	32.93	102.11
12	29.95	88.74
13	14.87	61.33
14	20.80	43.32
15	16.16	50.17
16	12.86	12.86
Average	**27.87**	**346.50**

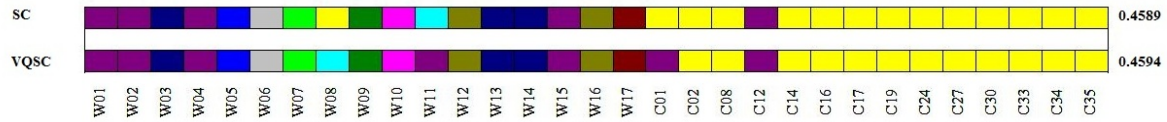


Figure 3.1: Cluster formation for SC and VQSC with Alignment Score and $m = 11$.

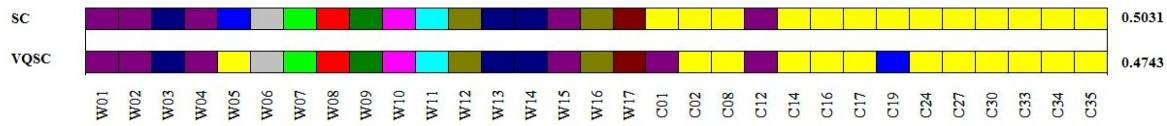


Figure 3.2: Cluster formation for SC and VQSC with Alignment Score and $m = 12$.

Table 3.5: Loss in cluster quality in terms of Silhouette Values because of sampling in SC.

# of Clusters	**% Loss in Silhouette Values**
2	0
3	+1.08
4	–6.19
5	–9.16
6	–27.58
7	–19.12
8	–36.52
9	–3.15
10	–19.81
11	–0.11
12	–6.07
13	–10.98
14	–9.04
15	–6.88
16	–18.25
Average	**-11.45**

As evident from Table 3.6, on an average only 4 out of 31 (about 13%) sequences are wrongly clustered by VQSC as compared with SC. This is considered acceptable because, as earlier, we are still better than the existing best algorithms (please see Table 3.7 and the accompanying discussion below).[3] To sum up, by using VQSC, we get almost the same cluster formation as SC, but at a reduced computational cost.

Finally, we compare results of our efficient and accurate algorithm (VQSC using Alignment Score) with the existing best (UPGMA and NJ). Results for this are given in Table 3.7. As evident from this table, our VQSC is on an average 21% better than

[3]The outlier case of $m = 8$ needs further analysis and experimentation with more data.

Table 3.6: Wrongly clustered sequences by VQSC when compared with SC.

# of Clusters	**# of Sequences Wrongly Clustered**
2	0
3	1
4	4
5	4
6	4
7	4
8	12
9	2
10	5
11	3
12	3
13	5
14	6
15	4
16	4
Average	**4.07**

UPGMA and on an average 24% better than NJ in terms of Silhouette Values. As earlier, we also have the added benefit of reduced computational complexity for VQSC as compared with both UPGMA and NJ.

Table 3.7: Comparison of VQSC with UPGMA and NJ.

# of Clusters	**% improvement in VQSC**	
	Over UPGMA	**Over NJ**
2	73.07	43.65
3	43.66	27.37
4	19.87	39.23
5	15.80	29.90
6	27.50	9.92
7	24.77	19.12
8	1.00	2.12
9	28.66	32.10
10	8.63	25.62
11	19.92	32.16
12	15.99	32.89
13	16.61	26.48
14	7.72	12.79
15	9.28	22.17
16	-2.63	7.27
Average	**20.66**	**24.19**

Chapter 4

Probabilistically Sampled Spectral Clustering for Phenotypic Data

The terms genotype and phenotype may sound similar but there is a considerable difference between them. A genotype refers to a plant variant defined by its Whole Genome Sequence (WGS), which is responsible for a particular trait or characteristics in the plant. For example, consider the following WGSs of an arbitrary plant, which map to different genotypes:

$$
\begin{array}{llllllllll}
\ldots & A & \boldsymbol{C} & G & T & \boldsymbol{G} & C & C & \boldsymbol{T} & A\ldots \quad \text{Genotype 1} \\
\ldots & A & \boldsymbol{A} & G & T & \boldsymbol{C} & C & C & \boldsymbol{A} & A\ldots \quad \text{Genotype 2} \\
\ldots & A & \boldsymbol{C} & G & T & \boldsymbol{G} & C & C & \boldsymbol{C} & A\ldots \quad \text{Genotype 3} \\
 & & & & & \vdots & & & & \\
\ldots & A & \boldsymbol{T} & G & T & \boldsymbol{G} & C & C & \boldsymbol{G} & A\ldots \quad \text{Genotype } n
\end{array}
$$

On the other hand, a phenotype refers to a plant variant defined by its physical characteristics. Let two similar plants may have different leaf colors or different heights. These differences between the color or the height might imply that these two plants are different phenotypes of the same genotype or they may belong to different genotypes altogether.

As mentioned above, phenotypic characteristics of a plant genotype refer to its physical properties as cataloged by plant biologists at different research centers around

the world. Clustering genotypes based upon their phenotypic characteristics is used to obtain diverse sets of parents that are useful in their breeding programs. The Hierarchical Clustering (HC) algorithm is the current standard in clustering of phenotypic data. This algorithm generates poor quality clusters and has high computational complexity. To address the cluster quality challenge, we propose the use of Spectral Clustering (SC) algorithm. To make the algorithm computationally cheap, we propose using sampling, specifically, Pivotal Sampling that is probability based. Since application of samplings to phenotypic data has not been explored much, for effective comparison, another sampling technique called Vector Quantization (VQ) is adapted for this data as well. Also, it has given promising results for genetic data as discussed in the previous chapter.

The novelty of our SC with Pivotal Sampling algorithm is in constructing the crucial similarity matrix for the clustering algorithm and defining probabilities for the sampling technique. Although our algorithm can be applied to any plant genotypes, we test it on the phenotypic data obtained from about 2400 Soybean genotypes. SC with Pivotal Sampling generates better quality clusters (in terms of Silhouette Values) than all the other proposed competitive clustering with sampling algorithms (i.e. SC with VQ, HC with Pivotal Sampling, and HC with VQ). The complexities of our SC with Pivotal Sampling algorithm and these three variants are almost same because of the involved sampling. In addition to this, SC with Pivotal Sampling outperforms the standard HC algorithm in both cluster quality and computational complexity. We experimentally show that we get 45% better quality clusters than HC in terms of Silhouette Values. The computational complexity of our algorithm is more than a magnitude lesser than HC.

The rest of this chapter is organized as follows. Section 4.1 provides a brief summary of the previous works on clustering of phenotypic data.[1] The crucial adaptations done in SC and Pivotal Sampling for phenotypic data are discussed in Section 4.2. Finally, Section 4.3 describes the experimental set-up, and the results.

[1]Since none of the previous works have used sampling for phenotypic data, we could not review this aspect.

4.1 Literature Review

In this section, we present some relevant previous studies on phenotypic data and the novelty of our approach. Broadly, these studies can be classified into two categories. The first category consists of the works that identify relationships between the different phenotypic characteristics (for example, lower plant height may relate to lower plant yield or vice versa). These works are discussed in Section 4.1.1. The second category consists of the studies that identify the genotypes having dissimilar phenotypic characteristics for the breeding program. These studies are discussed in Section 4.1.2. Finally, we present a set of works that belong to both the categories in Section 4.1.3.

4.1.1 First Category Previous Studies

Immanuel et al. [64] in 2011 measured nine characteristics of 21 Rice genotypes. Grain Yield (GY) was kept as the primary characteristic, and its correlations with all others were obtained. It was observed that characteristics like Plant Height (PH), Days to 50% Flowering (DF), Number of Tillers per Plant (NTP), Filled Grains per Panicle (FGP) and Panicle Length (PL) were positively correlated with GY. The remaining characteristics were negatively correlated with GY.

Divya et al. [65] in 2015 recorded 21 characteristics of two Rice genotypes. The authors investigated the association between Infected Leaf Area (ILA), Blast Disease Susceptibility (BDS), Number of Tillers per Plant (NTP), Grain Yield (GY) and others. The authors concluded that, for example, (a) ILA had a significant positive correlation with leaf's BDS, (b) NTP exhibited the highest association with GY.

Gireesh et al. [19] in 2015 analyzed eight characteristics of 3443 Soybean genotypes. The authors sampled these genotypes using two methods, and correlations of all the characteristics with each other for both the samples were estimated. It was observed that, for example, Days to 50% Flowering (DF) was positively correlated with Days to Pod Initiation (DPI) in both the samples, while Number of Pods Per Plant (NPPP) showed a negative correlation with Nodes Per Plant (NPP).

Huang et al. [66] in 2018 studied six characteristics of 206 Soybean genotypes. These characteristics were correlated with the three types of leaves; elliptical leaves, lanceolate leaves and round leaves. The authors deduced that Soybean plants with lanceolate leaves had maximum average Plant Height (PH), Number of Pods per Plant (NPP), Number of Branches per Plant (NBP), and 100-Seed Weight (SW), while Soybean plants with other two types of leaves had lower values of these characteristics.

Carpentieri-Pipolo et al. [67] in 2019 investigated 45 phenotypic characteristics of a Soybean genotype. The authors then studied the effect of 20 bacteria isolated from roots, leaves, and stems on these characteristics (i.e. whether the bacteria had positive or negative activity on (correlation with) the 45 characteristics). For example, *Enterobacter Ludwigii* (EL) bacteria, which is isolated from leaves, showed a positive correlation with 25 characteristics (e.g., Plant Growth Promotion (PGP)) and a negative correlation with remaining 20 characteristics (e.g., Phenylacetic Acid (PAC) assimilation). For better exposition, the above five studies are summarized in Table 4.1. Here, $\implies$ represents positive correlation and $\not\Longrightarrow$ represents negative correlation.

Table 4.1: Summary of first category previous studies.

Studies	**Plant**	**# of Genotypes**	**Inferred Relationship**
Immanuel et al. (2011)	Rice	21	PH, DF, NTP, FGP, PL $\implies$ GY
Divya et al. (2015)	Rice	2	ILA $\implies$ BDS
Gireesh et al. (2015)	Soybean	3443	DF $\implies$ DPI and NPPP $\not\Longrightarrow$ NPP
Huang et al. (2018)	Soybean	206	Lanceolate leaves $\implies$ max avg PH, NPP, NBP and SW
Carpentieri-Pipolo et al. (2019)	Soybean	1	EL $\implies$ PGP and EL $\not\Longrightarrow$ PAC

4.1.2 Second Category Previous Studies

Sharma et al. [14] in 2014 performed clustering of 24 synthetic Wheat genotypes (lines). Cluster analysis was performed using HC, and the genotypes were grouped into three clusters using the polymorphic Inter Simple Sequence Repeat (ISSR) markers. The authors argued that genotypes belonging to different clusters were diverse in terms of heat tolerance, and could be used to develop better heat tolerant genotype.

Kahraman et al. [15] in 2014 analyzed the field performance of 35 Common Bean genotypes by grouping them. The authors used HC, and the genotypes were clustered into three groups based upon the matrix of relationship between the genotypes. The genotypes belonging to different clusters were considered diverse, and were used to select promising genotypes for breeding.

Painkra et al. [10] in 2018 performed clustering of 273 Soybean genotypes. Here, the authors used HC, and the genotypes were grouped into seven clusters using Pearson Correlation Coefficient. According to the authors, the genotypes belonging to the distant clusters were more diverse such that choosing them maximized heterosis[2] in cross-breeding.

Islam et al. [68] in 2020 clustered ten Upland Rice genotypes. Here, HC was used and the genotypes were grouped into three clusters using a similarity coefficient between the genotypes. The authors identified the two best genotypes that could be used to obtain new genotypes having higher plant yield. As earlier, here also, we summarize the above four studies in Table 4.2 below.

4.1.3 Both Categories Previous Studies

Fried et al. [69] in 2018 analyzed 11 characteristics of 49 Soybean genotypes. The authors determined correlations between the root characteristics and other phenotypic characteristics. For example, Shoot Dry Weight (SDW) and Chlorophyll Index (CI) were positively correlated with Total Root Length (TRL) and Total Root Surface Area

[2]Heterosis refers to the phenomenon in which a hybrid plant exhibits superiority over its parents in terms of Plant Yield or any other characteristic.

Table 4.2: Summary of second category previous studies.

Studies	**Plant**	**# of Genotypes**	**Clustering Algorithm**	**# of Clusters**	**Development of Better Genotypes**
Sharma et al. (2014)	Wheat	24	HC	3	Heat Tolerant
Kahraman et al. (2014)	Common Bean	35	HC	3	Promising Genotypes for Breeding
Painkra et al. (2018)	Soybean	273	HC	7	Improved Characteristics
Islam et al. (2020)	Rice	10	HC	3	Higher Plant Yield

(TRSA), while Plant Height (PH) was negatively correlated with TRSA and Average Root Diameter (ARD). In this work, Principal Component Analysis (PCA) biplot was used to separate the genotypes into seven clusters. According to the authors, this research was critical for Soybean improvement programs since it helped select genotypes with the improved root characteristics.

Stansluos et al. [70] in 2019 analyzed 22 phenotypic characteristics for 11 Sweet Corn genotypes (cultivars). For example, the authors showed a positive and significant correlation of Yield of Marketable Ear (YME) with Ear Diameter (ED) and Number of Marketable Ear (NME), while a negative correlation between YME and Thousand Kernel Weight (TKW). Cluster analysis was performed using HC, and the corn genotypes were grouped into four clusters using the Ward Linkage. The authors inferred substantial variation in morphological and agronomic capabilities of different genotypes. Again, we summarize the above two studies in Table 4.3 below.

With the focus on the study of genetic diversity using phenotypic data, we have multiple novel contributions as below.

1. We focus on the second category above, and perform grouping of several thou-

Table 4.3: Summary of both categories previous studies.

Studies	**Plant**	**# of Genotypes**	**Inferred Relationship**	**Clustering Algorithm**	**# of Clusters**	**Development of Better Genotypes**
Fried et al. (2018)	Soybean	49	SDW, CI $\implies$ TRL, TRSA PH $\not\Rightarrow$ TRSA, ARD	PCA	7	Improved Root Characteristics
Stansluos et al. (2019)	Sweet Corn	11	YME $\implies$ ED, NME YME $\not\Rightarrow$ TKW	HC	4	Better Morphological Capabilities

sand genotypes as compared to a few hundred in the papers cited above. Note that from the first category, Gireesh et al. [19] did work with about three thousand genotypes, and we do compare one aspect of our work with this previous work (more on this in the point 2a below).

2. Clustering becomes computationally expensive when the size of the data is very large. Hence, sampling is required to make the underlying algorithm scalable. Thus, we perform clustering on the sampled data rather than the full one, which is not done in any of the papers above. We have two more innovations in this aspect as below.

   (a) We use a probability-based sampling technique (Pivotal Sampling as mentioned earlier) that is highly accurate, and forms a completely new contribution. We demonstrate the superiority of our sampling by comparing it with the one done in Gireesh et al. [19]. This comparison is discussed towards the end of the Results section. Please note that Gireesh et al. only performed sampling and did not cluster their data.

   (b) HC, which is the most common clustering algorithm (and some other sporadically used algorithms like k-means and UPGMA), do not generate the good quality clusters. Again, as earlier, we develop a variant of the SC algorithm, which is considered to give high quality clusters, especially for phenotypic data. Use of SC in this context is also completely new. We show the dominance of our clustering algorithm over the one proposed in

the most recent past work by Islam et al. [68] towards the end of the Results section. Again, please note that Islam et al. only performed clustering and did not sample their data.

4.2 Our Algorithm

Consider that the phenotypic data of a plant consists of n genotypes with each genotype evaluated for m different characteristics/ traits. These characteristics may have categorical (non-numerical) or numerical values. Hence, we need to convert the categorical values into numerical ones. For this, we use the label encoder method [71]. This method transforms non-numerical labels into numerical values between 0 and (number of categories) – 1. For example, if a characteristic has three possible labels; poor, good, and very good, we use 0, 1, and 2 to represent them, respectively.

Since different characteristics have values in different ranges, we start by normalizing them as below [72, 73]

$$(\mathfrak{X}_j)_i = \frac{(x_j)_i - \min(x_j)}{\max(x_j) - \min(x_j)}. \tag{4.1}$$

Here, $(\mathfrak{X}_j)_i$ and $(x_j)_i$ are the normalized value and the actual value of the j^{th} characteristic for the i^{th} genotype, respectively with $j = 1, ..., m$ and $i = 1, ..., n$. Furthermore, $\max(x_j)$ and $\min(x_j)$ are the maximum and the minimum values of the j^{th} characteristic among all the genotypes. We use this normalized data for clustering and sampling. Now, we give the implementation of our modified SC algorithm on phenotypic data. Subsequently, we present the application of Pivotal Sampling to obtain the samples from the same data.

4.2.1 Implementing Modified SC for Phenotypic Data

Similar to the standard SC algorithm discussed in Section 2.1, the first step in our modified SC is to obtain the similarity matrix. Let vector p_i contain the normalized values of all the characteristics (m) for the i^{th} genotype. We define the similarity between the vectors p_1 and p_2 (without loss of generality, representing the genotypes

1 and 2, respectively) as the inverse of the distance between these vectors obtained by using the seven different distance measures: Euclidean, Squared Euclidean, City-block, Cosine, Correlation, Hamming and Jaccard [50]. This is intuitive because smaller the distance between any two genotypes, larger the similarity between them and vice versa. We denote this distance by $d_{p_1p_2}$ and build the similarity matrix by obtaining the similarities among all the genotypes.

The next step is to compute the Laplacian matrix, which when obtained from the above-discussed similarity matrix, generates poor eigenvalues,[3] and in turn poor corresponding eigenvectors that are required for clustering.[4] Thus, instead of taking only the inverse of $d_{p_1p_2}$, we also take its exponent, i.e. we define the similarity between the genotypes 1 and 2 as $e^{-d_{p_1p_2}}$ [4]. This, besides fixing the poor eigenvalues/ eigenvectors problem, also helps perform better clustering of the given data. Further, we follow the remaining steps of standard SC as given in Section 2.1.3.

4.2.2 Applying Pivotal Sampling to Phenotypic Data

Pivotal Sampling requires that the inclusion probabilities (i.e. π_i for $i = 1, ..., n$), of all the units (genotypes here) in the population U, be computed before a unit is considered for a contest. The set of characteristics associated with a genotype can be exploited in computing these probabilities. As mentioned in Section 2.3, to obtain a sample of size N, where $N \ll n$, we calculate these probabilities as

$$\pi_i = N \frac{\varkappa_i}{\sum_{i \in U} \varkappa_i}, \tag{4.2}$$

where $\varkappa_i$ can be a property associated with the data.

In our implementation, we use the deviation property of the genotypes to obtain $\varkappa_i$, which for the i^{th} genotype is calculated using the normalized values as

$$dev_i = \sum_{j=1}^{m} \max(\mathfrak{X}_j) - (\mathfrak{X}_j)_i,$$

[3]Zero/ close to zero and distinct eigenvalues are considered to be a good indicator of the connected components in a similarity matrix. Thus, eigenvalues are considered poor when they are not zero/ not close to zero or indistinct.

[4]For some distance matrices (like Euclidean distance), the eigenvalues don't even converge.

where $\max(\mathfrak{X}_j)$ denotes the maximum normalized value of the j^{th} characteristic among all the genotypes and $(\mathfrak{X}_j)_i$ is given by (4.1). Practically, a relatively large value of dev_i indicates that the i^{th} genotype is less important, and hence, its probability should be small. Thus, the inclusion probability of a genotype is calculated by taking $\varkappa_i = \frac{1}{dev_i}$ in (4.2) or

$$\pi_i = N \frac{\frac{1}{dev_i}}{\sum_{i \in U} \frac{1}{dev_i}}.$$

Once these probabilities are obtained, we follow the two steps (selection and rejection) as discussed in Section 2.3 to obtain N samples. These N sampled genotypes are then grouped into k clusters using our modified SC discussed in the previous subsection.

However, our goal is to cluster all n genotypes and not just N. Hence, there is a need to reverse-map the remaining $n - N$ genotypes to these k clusters. For this, we define the notion of average similarity, which between the non-clustered genotype $\tilde{p}$ and the cluster C_l is given as

$$\mathcal{AS}(C_l, \tilde{p}) = \frac{1}{\#(C_l)} \sum_{q \in C_l} e^{-d_{\tilde{p}q}}.$$

Here, $\#(C_l)$ denotes the number of genotypes present in C_l and q is a genotype originally clustered in C_l by our modified SC algorithm with Pivotal Sampling. As earlier, $d_{\tilde{p}q}$ denotes the distance between the genotypes $\tilde{p}$ and q. We obtain the average similarity of $\tilde{p}$ with all the k clusters (i.e. with C_l for $l = 1, ..., k$), and associate it with the cluster with which $\tilde{p}$ has the maximum similarity.

Next, we perform the complexity analysis of our algorithm. Since SC and Pivotal Sampling form the bases of our algorithm, we discuss the complexities of these algorithms before ours.

1. SC (n: number of genotypes, m: number of characteristics)
   (a) Constructing Similarity Matrix: $\mathcal{O}(n^2 m)$
   (b) Obtaining Laplacian Matrix: $\mathcal{O}(n^3)$

2. Pivotal Sampling (n, N: sample size)
   (a) Obtaining Probabilities: $\mathcal{O}(n)$
   (b) Obtaining Samples: $\mathcal{O}(n)$

3. Our Algorithm (n, m, N)
   (a) Obtaining Samples: $\mathcal{O}(n)$
   (b) Constructing Similarity Matrix: $\mathcal{O}(N^2m)$
   (c) Obtaining Laplacian Matrix: $\mathcal{O}(N^3)$
   (d) Reverse Mapping: $\mathcal{O}\big((n-N)N\big)$

Thus, the overall complexity of our algorithm is $\mathcal{O}(nN + N^3 + N^2m)$. Here, we have kept three terms because any of these can dominate (here, $n \gg N, m$). When we compare complexity of our algorithm with that of HC, which is $\mathcal{O}(n^3)$, it is evident that we are more than a magnitude faster than HC. We revisit this complexity analysis after discussing data in the next section, which supports our claim further.

4.3 Results

In this section, we first briefly discuss the data used for our experiments. Next, we describe the clustering set-up, where the ideal number of clusters, the suitable distance measures for building similarity matrices, and the most useful Laplacian matrix are discussed. Subsequently, we present the results for our modified SC with Pivotal Sampling. Here, we compare our algorithm with (a) SC with VQ, HC with Pivotal Sampling, HC with VQ and (b) non-sampled HC. Finally, we give the goodness of our sampling technique by estimating a measure called the population total.

4.3.1 Data Description

As mentioned earlier, our techniques can be applied to any plant data. However, here we experiment on phenotypic data of Soybean genotypes. This data is taken from Indian Institute of Soybean Research, Indore, India, and consists of 29 different characteristics/ traits for 2376 Soybean genotypes [19]. Among these, we consider the following eight characteristics that are most important for higher yield: Early Plant Vigor (EPV), Plant Height (PH), Number of Primary Branches (NPB), Lodging Score (LS), Number of Pods Per Plant (NPPP), 100 Seed Weight (SW), Seed Yield Per Plant (SYPP) and Days to Pod Initiation (DPI). Out of these, EPV and LS have

categorical values, while the remaining characteristics have numerical values. Hence, we convert these two categorical values into numerical ones using the label encoder method discussed in the previous section. A snapshot of this phenotypic data for a few Soybean genotypes is given in Appendix A. Here, we also perform validation of this data by comparing it with a similar dataset.

Next, we compare the complexities of our algorithm and HC using the selected data; see Table 4.4. It is evident from this table that our algorithm achieves substantial savings.

Table 4.4: Computational complexity comparison for the given data.

# of Genotypes (n)	**# of Characte-ristics** (m)	**Sample Size** (N)	**Our Algorithm** ($nN + N^3 + N^2m$)	**HC** (n^3)
2376	8	500	$(2376 \times 500) + (500)^3 + (500)^2 \times 8$ $= 1.28 \times 10^8$	$(2376)^3$ $= 1.34 \times 10^{10}$
2376	8	300	$(2376 \times 300) + (300)^3 + (300)^2 \times 8$ $= 2.84 \times 10^7$	$(2376)^3$ $= 1.34 \times 10^{10}$

4.3.2 Clustering Setup

Here, *first*, we determine the ideal number of clusters by using the eigenvalue gap heuristic [17, 74]. If $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of the matrix used for clustering (e.g., the Laplacian matrix), then often the initial set of eigenvalues, say k, have a considerable difference between the consecutive ones in this set. That is, $|\lambda_i - \lambda_{i+1}| \not\approx 0$ for $i = 1, ..., k-1$. After the k^{th} eigenvalue, this difference is usually approximately zero. According to this heuristic, this k gives a good estimate of the ideal number of clusters.

For this experiment, without loss of generality, we build the similarity matrix using the Euclidean distance measure on the above discussed phenotypic data. As mentioned in Section 2.1.2, it is recommended to use the random walk Laplacian matrix (L_{rw})

[17]. Hence, we use its eigenvalues for estimating k. Figure 4.1 represents the graph of the first fifty smallest eigenvalues (in absolute terms) of this Laplacian matrix. On the x-axis, we have the eigenvalue number, and on the y-axis its corresponding value.

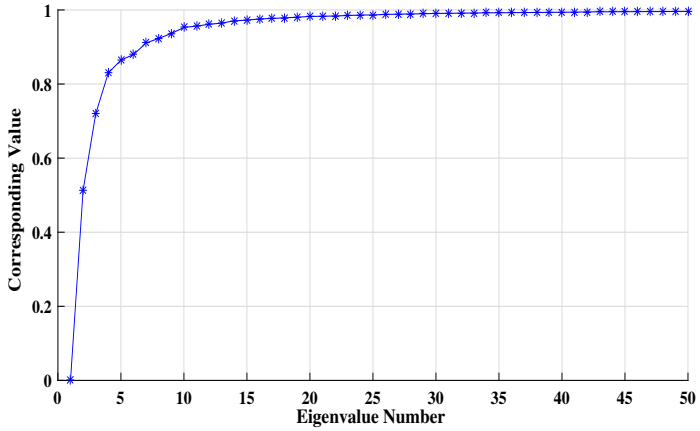


Figure 4.1: Fifty Smallest Eigenvalues of the Type-3 Laplacian Matrix Obtained from the Euclidean Similarity Matrix (for estimating the ideal number of clusters).

From this figure, we can see that there is a considerable difference between the first ten consecutive eigenvalues. After the tenth eigenvalue, this difference is very small (tending to zero). Hence, based upon the earlier argument and this plot, we take k as ten. To corroborate this choice more, we experiment with k as twenty and thirty as well. As expected, and discussed in detail later in this section, Silhouette Values for these numbers of clusters are substantially lower than those for ten clusters.

Second, and final, we perform experiments to identify the suitable similarity measures to build the similarity matrix, and also verify that, as recommended, the L_{rw} Laplacian matrix is the best. For this work as well, we use Silhouette Value as the validation metric, which is discussed in Section 3.3.2. Table 4.5 below gives Silhouette Values of our modified SC for all seven similarity measures[5] and three Laplacians[6] when clustering the earlier presented phenotypic data into 10, 20, and 30 clusters.

From this table, it is evident that Silhouette Values for the Euclidean, Squared Eu-

[5]We use Euclidean, Squared Euclidean, City-block, Cosine, Correlation, Hamming, and Jaccard similarity measures [50].

[6]One non-normalized Laplacian matrix (L) and two normalized Laplacian matrices (L_{sym} & L_{rw}).

Table 4.5: Silhouette Values for modified SC with seven similarity measures and three Laplacian matrices for $k = 10, 20$, and 30. Silhouette Values in bold represent good clustering.

Sr. No.	**Similarity Measure**	**# of Clusters (k)**	$\boldsymbol{L}$	$\boldsymbol{L_{sym}}$	$\boldsymbol{L_{rw}}$
1.	Euclidean	10	0.0828	-0.0273	**0.2422**
		20	0.0455	-0.1096	**0.2069**
		30	0.0887	-0.1536	**0.1783**
2.	Squared Euclidean	10	0.0815	-0.0555	**0.3836**
		20	-0.0315	-0.1809	**0.2612**
		30	0.0354	-0.2367	**0.1538**
3.	City-block	10	0.0687	0.2375	**0.2647**
		20	-0.0356	0.1347	**0.2082**
		30	-0.0870	0.0866	**0.1887**
4.	Cosine	10	0.1737	-0.1408	0.0694
		20	0.0359	-0.1973	0.0277
		30	0.0245	-0.2456	-0.0316
5.	Correlation	10	0.1926	-0.1259	**0.3426**
		20	0.0970	-0.2198	**0.2313**
		30	0.2383	-0.2604	**0.1556**
6.	Hamming	10	0.0643	0.0706	0.0775
		20	0.0683	0.0311	0.0382
		30	0.0715	0.0283	0.0229
7.	Jaccard	10	0.0716	0.0303	0.0458
		20	0.0446	0.0276	0.0236
		30	0.0279	0.0298	0.0318

clidean, City-block and Correlation similarity measures and the L_{rw} Laplacian matrix are the best. Hence, we use these four similarity measures and this Laplacian matrix. Also, as mentioned earlier, Silhouette Values decrease for twenty and thirty cluster sizes.

4.3.3 Clustering and Sampling Results

Using the earlier presented dataset, and clustering-sampling setups, we compare our proposed algorithm (modified SC with Pivotal Sampling) with the existing variants in four ways. Again, as earlier, we use Silhouette Values for comparison. Quantifying statistical difference between different Silhouette Values is a hard task. In general, the more closer these values are to one, the better is the clustering (see Section 3.3.2).

First, we demonstrate that use of sampling with modified SC does not deteriorate the quality of clustering. Second, we compare our algorithm with modified SC with VQ, HC[7] with Pivotal Sampling and HC with VQ for a sample size of 500. Since the results for modified SC with VQ come out to be closest to our algorithm, next, for broader appeal we compare these two algorithms for a sample size of 300. Third, we compare our algorithm with the current best in literature for this kind of data (i.e. HC without sampling) for both the sample sizes of 500 and 300. Fourth and finally, as discussed in the Literature Review section, we compare our sampling with that in Gireesh et al. [19] and our clustering with the one in Islam et al. [68].

Initially, we calculate the loss in cluster quality incurred in terms of Silhouette Values because of Pivotal Sampling in our algorithm. This loss for both the sample sizes and cluster size ten is listed in Table 4.6. Columns 1 and 2 give the sample sizes and the similarity measures chosen, respectively. Columns 3 and 4 give the Silhouette Values for modified SC without sampling (from Table 4.5) and our algorithm, respectively. The last column gives the percentage loss in Silhouette Values. We can observe from this data that the loss for one type of similarity measure (Correlation) is almost as low as -2% for both the sample sizes. This is considered acceptable because we are still better than the existing best algorithm (HC without sampling; please see Table 4.9 and its accompanying discussion below).

Here, we also perform a statistical test to support the above conjecture that using Pivotal Sampling does not substantially deteriorate the quality of clusters obtained by our modified SC. For this, we use the ANOVA (analysis of variance) test [75].

[7]HC also requires building a similarity matrix.

Table 4.6: Loss in Silhouette Values because of Pivotal Sampling in modified SC for cluster size ten.

Sample Size	**Similarity Measure**	**modified SC**	**modified SC with Pivotal Sampling**	**Percentage Loss in Silhouette Value**
$N = 500$	Euclidean	0.2422	0.2152	-11.15%
	Squared Euclidean	0.3836	0.3362	-12.36%
	City-block	0.2647	0.2369	-10.50%
	Correlation	0.3426	0.3367	-1.72%
$N = 300$	Euclidean	0.2422	0.2104	-13.13%
	Squared Euclidean	0.3836	0.3280	-14.49%
	City-block	0.2647	0.2392	-9.63%
	Correlation	0.3426	0.3368	-1.69%

This test uses the variance between the different groups and the variance within each group to compute a value called the F-value, which is then compared with a standard estimate called F-critical. If F-value is less than F-critical, then it is inferred that the means of all the groups are equal.

The two groups for us refer to the modified SC results (column 3) and the modified SC with Pivotal Sampling results (column 4). The F-values here (using the Silhouette Values of the two groups) come out to be 0.3432 and 0.4202 for N = 500 and N = 300, respectively. Both these values are less than the F-critical value given in the F-distribution table of [76], which is 5.9873. Thus, using the above mentioned ANOVA test theory, we infer that that the mean Silhouette Value of modified SC is similar to the mean Silhouette Value of modified SC with Pivotal Sampling for both the sample sizes.

The results for the *second* set of comparisons are given in Table 4.7. Columns 2 and 3 give the similarity measures and the number of clusters chosen, respectively. Columns 4 and 5 give Silhouette Values of modified SC with Pivotal Sampling and VQ, respectively, while columns 6 and 7 give Silhouette Values of HC with Pivotal

Sampling and VQ, respectively.

Table 4.7: Silhouette Values for modified SC and HC with Pivotal Sampling and VQ for $N = 500$. Silhouette Values in bold represent good clustering.

Sr. No.	**Similarity Measure**	**# of Clusters (k)**	**modified SC**		**HC**	
			Pivotal Sampling	**VQ**	**Pivotal Sampling**	**VQ**
1.	Euclidean	10	**0.2152**	0.2061	0.2105	-0.1040
		20	**0.1905**	0.1448	0.2263*	-0.1620
		30	**0.1741**	0.1021	0.1933*	-0.2874
2.	Squared Euclidean	10	**0.3362**	0.2969	0.2634	-0.2096
		20	**0.2469**	0.1522	0.3726*	-0.5899
		30	**0.1658**	0.0440	0.2933*	-0.6083
3.	City-block	10	**0.2369**	0.2354	0.1703	-0.2278
		20	**0.2019**	0.1870	0.1879	-0.2398
		30	**0.1752**	0.1524	0.1988*	-0.2868
4.	Correlation	10	**0.3367**	0.2560	0.2582	-0.0060
		20	**0.2291**	0.0899	0.0867	-0.4120
		30	**0.1742**	-0.0349	0.0998	-0.7018

When we compare our algorithm (values in the fourth column, and highlighted in bold) with other variants, it is evident that we are clearly better than modified SC with VQ and HC with VQ (values in the fifth and the seventh columns) as our values are higher than those from these two algorithms.

When we compare our algorithm with HC with Pivotal Sampling (values in the sixth column), we again perform better for many cases. However, for some cases, our algorithm performs worse than HC with Pivotal Sampling (highlighted with a *). Upon further analysis (discussed below), we realize that segregation of genotypes by HC with Pivotal Sampling into fewer clusters than practically observed, results in these set of Silhouette Values getting wrongly inflated.

To further assess the quality of the proposed technique, we present the distribution of genotypes into different clusters (after reverse-mapping) for HC with Pivotal Sampling and our algorithm. Without loss of generality, this comparison is done using the

Squared Euclidean similarity measure and cluster size ten. The results for HC with Pivotal Sampling are given in Figure 4.2 and for our algorithm are given in Figure 4.3. In both the figures, on the x-axis, we have the cluster number and on the y-axis, the number of genotypes present in them.

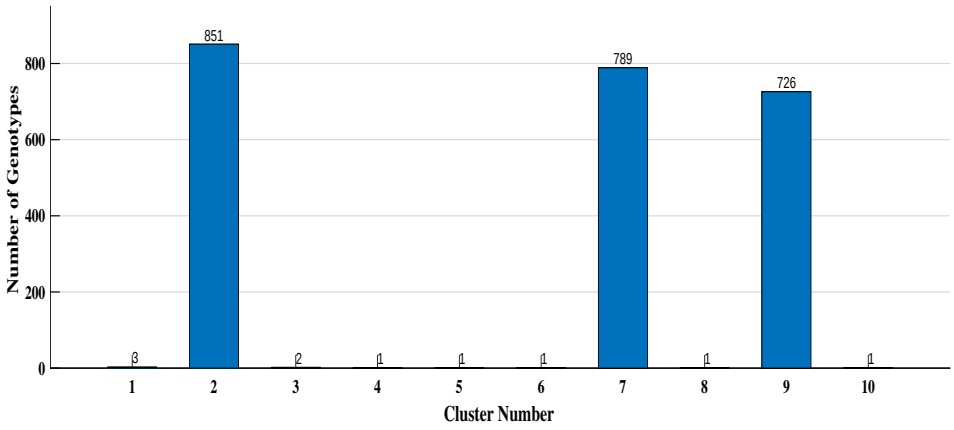


Figure 4.2: Distribution of Genotypes (HC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size ten.

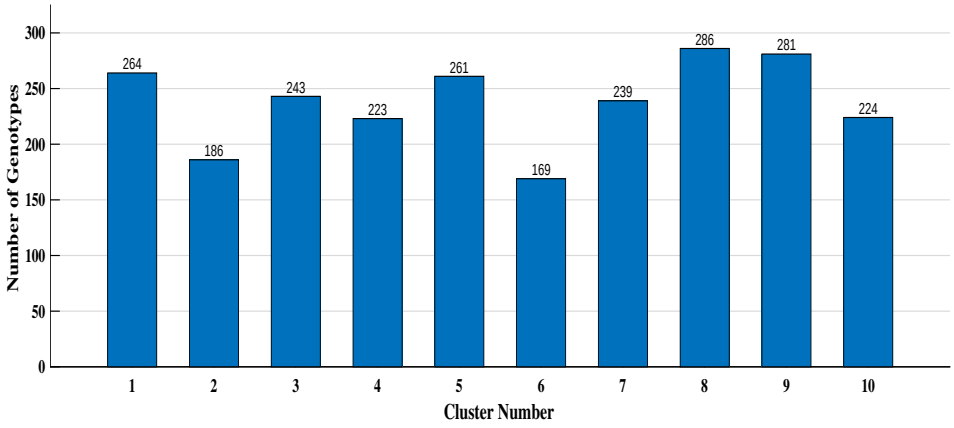


Figure 4.3: Distribution of Genotypes (modified SC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size ten.

As evident, Figure 4.2 depicts a very skewed distribution, i.e. most genotypes are segregated into only a few clusters, while the remaining clusters contain only one or two genotypes. At a broader level, this biased distribution of genotypes obtained by

HC with Pivotal Sampling is correct since all genotypes belong to the same plant. On the contrary, the distribution in Figure 4.3 is fairly equal. That is, our algorithm equally distributes all genotypes between the different clusters. At a finer level, this distribution is better since our algorithm is able to perform a more detailed clustering, i.e. it splits the bigger clusters into multiple smaller ones, which better captures the similarity between genotypes.

This is also the reason for the inflation of Silhouette Values of HC with Pivotal Sampling in Table 4.7 since the intra-cluster similarity for solitary genotype is zero leading to its respective Silhouette Value to become one (the maximum possible; see (3.1)). Thus, our algorithm also outperforms HC with Pivotal Sampling, which from Table 4.7 was not very evident.

Next, as mentioned earlier, to further demonstrate the applicability of our work, we also present the results with a sample size 300. Since modified SC with VQ turns out to be our closest competitor, we compare our algorithm with this one only. This comparison is given in Table 4.8, with its columns mapping the respective columns of Table 4.7. As evident from Table 4.8, our modified SC with Pivotal Sampling substantially outperforms modified SC with VQ (see values in columns 4 and 5).

As earlier, *third*, we compare the results of our algorithm (modified SC with Pivotal Sampling) with the currently popular clustering algorithm in the plant studies domain (i.e. HC without sampling). For this set of experiments, without loss of generality, we use the cluster size of ten. The results of this comparison are given in Table 4.9, where the first four columns are self-explanatory (based upon the data given in Tables 4.7 and 4.8 earlier). In the last column of this table, we also evaluate the percentage improvement in our algorithm over HC. As evident from this table, our algorithm generates 45% better quality clusters than HC in terms of Silhouette Values for both the sample sizes. As earlier, our algorithm also has the crucial added benefit of reduced computational complexity as compared to HC.

Fourth and *finally*, as mentioned in the Literature Review section, we also compare our work with two previous works that are closest to ours. With the dataset almost the same as used by us, that is, a slightly larger phenotypic data for Soybean genotypes,

Table 4.8: Silhouette Values for modified SC with Pivotal Sampling and VQ for $N = 300$.

Sr. No.	**Similarity Measure**	**# of Clusters (k)**	**modified SC**	
			Pivotal Sampling	**VQ**
1.	Euclidean	10	0.2104	0.1833
		20	0.1968	0.0955
		30	0.1743	0.0722
2.	Squared Euclidean	10	0.3280	0.2589
		20	0.2424	0.1322
		30	0.1613	0.0044
3.	City-block	10	0.2392	0.2157
		20	0.1990	0.1696
		30	0.1752	0.1373
4.	Correlation	10	0.3368	0.2229
		20	0.2312	0.0336
		30	0.1725	-0.0788

Table 4.9: Silhouette Values of modified SC with Pivotal Sampling and HC for cluster size ten.

Sample Size	**Similarity Measure**	**modified SC with Pivotal Sampling**	**HC**	**Percentage Improvement**
$N = 500$	Euclidean	0.2152	0.2173	-0.97%
	Squared Euclidean	0.3362	0.3257	3.22%
	City-block	0.2369	0.2135	10.96%
	Correlation	0.3367	0.2307	45.95%
$N = 300$	Euclidean	0.2104	0.2173	-3.28%
	Squared Euclidean	0.3280	0.3257	0.71%
	City-block	0.2392	0.2135	12.04%
	Correlation	0.3368	0.2307	45.99%

Gireesh et al. [19] performed Principal Component and Power Core based samplings to identify relationships between the different phenotypic characteristics (from Section 4.1.1). We compare our sampling results with the best from [19] in Appendix B, which

demonstrates the superiority of our sampling method. Islam et al. [68] performed HC on phenotypic data for Rice genotypes (from Section 4.1.2). In Appendix C, we apply modified SC on this dataset to again demonstrate that our clustering technique is better.

4.3.4 Sampling Estimators

To inspect the quality of our sampling techniques, we estimate a measure called the population total, which is the addition of values of a particular characteristic for all the n units (genotypes here) present in the population U. For example, if "Plant Height (PH)" is the characteristic of interest, then the population total is the addition of PH values for all the n genotypes. Mathematically, the exact (or actual) population total for a characteristic of interest x_j is given as

$$Y = \sum_{i \in U} (x_j)_i, \tag{4.3}$$

where, as earlier, $(x_j)_i$ is the value of the j^{th} characteristic for the i^{th} genotype and U is the set of all genotypes. By the definition of this measure (and also for two more measures listed below), we work with original (non-normalized) values of the characteristics rather than normalized ones. Also, based upon the same argument, we work with only those characteristics that are originally numerical.

In this work, we use two different estimators to compute an approximation of the population total from the sampled data. Closer the value of an estimator to the actual value, better the sampling. First is the Horvitz-Thompson (HT)-estimator (also called π-estimator), which is defined as [77]

$$Y'_{HT} = Y'_{\pi} = \sum_{i \in S} \frac{(x_j)_i}{\pi_i}, \tag{4.4}$$

where, π_i is the inclusion probability of the i^{th} genotype as evaluated in Section 4.2.2 and S is the set of sampled genotypes. Another estimator that we use is the Hájek-estimator. It is usually considered better than the HT-estimator and is given as [78]

$$Y'_{Hájek} = n \frac{\sum_{i \in S} \frac{(x_j)_i}{\pi_i}}{\sum_{i \in S} \frac{1}{\pi_i}}, \tag{4.5}$$

here, as earlier, n is the total number of genotypes.

The actual population total and the values of the above two estimators for six characteristics (that have numerical values) when using Pivotal Sampling and 500 samples are given in Table 4.10 (see columns 3, 4, and 6, respectively). From this table, it is evident that the approximate values of the population total are very close to the corresponding actual values. Thus, Pivotal Sampling works well in an absolute sense. Here, we also compute the values of the two estimators when using VQ (see columns 5 and 7). We can notice from these results that VQ also works reasonably well, but Pivotal Sampling is better.

Table 4.10: HT and Hájek estimators values for Pivotal Sampling and VQ as compared to the actual population total with $N = 500$ as the sample size.

Characteristics	**Actual Population Total**	**Pivotal Sampling (HT)**	**VQ (HT)**	**Pivotal Sampling (Hájek)**	**VQ (Hájek)**
PH	121773.05	122507.84	123407.80	123716.09	113168.90
NPB	8576.56	8585.28	9669.29	8669.95	8867.05
NPPP	99712.72	100193.53	114465.66	101181.70	104968.67
SW	20073.32	19907.10	20966.86	20103.44	19227.28
SYPP	10048.04	10137.57	10536.08	10237.55	9661.92
DPI	136810	135309.78	149242.17	136644.29	136859.84

Chapter 5

Classification of the Mammogram Patches

Breast cancer is becoming pervasive with each passing day. Hence, its early detection is a big step in saving the life of any patient. Mammography is a common tool in breast cancer diagnosis. The most important step here is classification of mammogram patches as normal–abnormal and benign–malignant.

As mentioned earlier, texture of a breast in a mammogram patch plays a significant role in these classifications. Hence, we propose a variation of Histogram of Gradients (HOG) [31] and Gabor filter [79] combination called Histogram of Oriented Texture (HOT) that exploits this fact. We also revisit the Pass Band - Discrete Cosine Transform (PB-DCT) descriptor that captures texture information well. All features of a mammogram patch may not be useful. So, we apply a feature selection technique called Discrimination Potentiality (DP). Support Vector Machine (SVM) is the most suitable classifier for two-class classification and is widely used in this field. Hence, we use this.

Density of a mammogram patch is important for classification, and has not been studied exhaustively. The Image Retrieval in Medical Application (IRMA) database [36] from RWTH Aachen, Germany is a standard database that provides mammogram patches, and most researchers have tested their frameworks only on a subset of patches from this database. We apply our *two* descriptors (DP-HOT and DP-PB-DCT) on

all images of the IRMA database for density-wise classification, and compare with the standard descriptors. We achieve higher accuracy than all of the existing standard descriptors (more than 92%).

The rest of this chapter is organized as follows. Section 5.1 provides the summary of the related work. The proposed mammogram patch classification system is explained in Section 5.2. Finally, Section 5.3 presents the experimental results.

5.1 Literature Review

Researchers have reviewed existing techniques for detection and analysis of abnormalities in mammogram patches like calcification, masses, tumors, bilateral asymmetry, and architectural distortion, etc. [22]. Some people have also reviewed the contribution of texture to risk assessment for each density separately [21]. Mammogram patches consist of directionally oriented, texture image due to its fibro-glandular tissues, ligaments, blood vessels and ducts. These texture features for mammogram patches can be categorized into four groups; statistical [80, 34, 81, 82, 83] local pattern histogram [84, 32, 31, 35], directional [85, 86, 87], and transform-based [88, 89].

Statistical features such as mean, variance, energy, entropy, skewness, and kurtosis are mostly utilized as a descriptor for classification [80, 34, 81, 82, 83]. Gray Level Co-occurrence Matrix (GLCM) and Gray Level Run Length Matrix (GLRLM) provide the relationship between neighboring pixels of a mammogram patch. Statistical properties of these matrices have also been exploited for mammogram patch classification. These features are extracted by directly using spatial data from images.

Some works have exploited local distribution of textural properties of mammogram patches for classification. HOG [31], Local Configure Pattern (LCP) [31], Uniform Directional Pattern (UDP) [84], Local Ternary Pattern (LTP) [90], Local Phase Quantization (LPQ) [91] and Local Binary Pattern (LBP) [92] are some such examples. There are three different variants of LBP, which are usually used for exploiting local textural properties; Uniform Local Binary Pattern (LBP-u), Rotation Invariant Local Binary Pattern (LBP-ri) and Rotation Invariant Uniform Local Binary Pattern

(LBP-riu) [93]. Block-wise feature extraction gives better performance as compared to global feature vectors. Statistical properties of local histogram have also been used as a mammogram patch descriptor for classification [94].

Coming to directional features, wavelet, dual-tree complex wavelet Gabor, Contourlet, finite Shearlet, etc. have been exploited for multi-resolution and multi-orientation texture or tissue analysis of a mammogram patch [86, 87, 26, 95]. Gabor based feature extraction schemes are widely used for mass classification as benign–malignant. Gabor features can be extracted from mammogram patches in different ways [85, 96]. Recently, some people have proposed directional features of mammogram patches computed by a Gabor wavelet for four different scales and eight different orientations [85]. Optimal parameters of a Gabor filter increases discrimination between normal and abnormal properties.

Finally, the fourth category for texture feature extraction is by using a mathematical transform. For example, Discrete Cosine Transform is one such option [88, 89].

In this work, we first propose a descriptor that exploits local distribution of textural property (HOG) as well as considers directional features (Gabor). We term it as HOT. *The reason for deriving this descriptor is that, for density-based mammogram patch classification, individually these two descriptors have their own drawbacks, which are eliminated in their combination.* The width of tissues may vary with the density of a mammogram patch, and it is difficult to estimate with HOG. Applying a Gabor filter for feature extraction on the whole mammogram patch is not useful since abnormalities are usually very local. The combination of HOG and Gabor filter is not new (see [97, 98, 99]). However, none of these works have applied this combination to mammogram patch classification. Moreover, our combination is optimally designed for solving the problem at-hand. We do a detailed comparison of our descriptor with these existing ones at the end of Section 5.2.2.1, i.e. after describing our descriptor.

The HOT descriptor mentioned above has few drawbacks in terms of capturing all textural features. Next, we revisit a transform based descriptor; Pass Band - Discrete Cosine Transform (PB-DCT). This descriptor has not been used yet for density-based mammogram patch classification. DCT has very strong energy compaction capability,

i.e., an image can be represented by a small set of coefficients. DCT coefficients are divided into three bands; high, middle and low. The high-frequency coefficients correspond to irrelevant information, the medium frequency coefficients carry textural information, and the low-frequency coefficients contain illumination information. We use PB-DCT like a band pass filter to extract mostly the middle frequencies with some amount of low frequencies as well.

The dimension of features can be reduced by either using a feature selection scheme or a dimension reduction scheme [86]. Feature selection schemes select the appropriate feature set based upon a criterion (such as entropy, fisher, maximum mutual information, etc.), while dimension reduction schemes project features onto an other dimensional subspace using orthogonal matrices. If the number of training samples for each class is less than the dimension of features, it is known as small sample size (SSS). Under this circumstance, which is common, some matrices in the dimension reduction approach become singular leading to difficulty in further computation [86]. In general, it has been found from literature that the rank-based feature selection approach is more suitable for feature reduction. Hence, we use this. Genetic algorithm can also be utilized for selecting suitable features from intensity, texture and shape features for benign and malignant classification of mammogram patches. This is part of future work.

Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Fisher Linear Discriminant (FLD), Naive Bayes and Neural Network with Multi-Layer Perception Learning have been utilized for mammogram patch classification [86, 80, 94]. SVM is the most suitable classifier for two class classification and is widely used in this field. Hence, we use this. Recently, ensembles of two or more classifiers (also called as multiclassifiers) have been used to improve the classification accuracy (see [93]). In this work, first, different descriptors are obtained by varying certain parameters. Then, after extracting features by using each of these descriptors, different classifiers are trained. Finally, these classifiers are combined by some technique (e.g., different SVMs are combined by a sum rule). This type of work will be explored in future.

Mammogram patches are usually categorized into four classes based upon the level

of density (i.e. fat transparent (d), fibro-glandular (e), heterogeneously dense (f), and extremely dense (g)). This is called the Breast Imaging Reporting And Database Systems (BIRADS) classification. Each BIRADS category is divided into three classes as normal, benign and malignant [100, 36, 101].

The IRMA reference database is a repository of mammogram patches and has been created by Deserno et al. [36] to test the accuracy of approaches for mammogram patch classification. It contains the two datasets, MIAS and DDSM. This database provides information about images based on the type of background tissue and the class of abnormality present in the mammogram patch. Table 5.1 lists the number of images in both the MIAS and DDSM datasets based upon the above discussed classification. Some sample patches are shown in Fig. 5.1

Table 5.1: Distribution of normal, benign, and malignant mammogram patches of the two different datasets for the four BIRADS classes.

IRMA: MIAS Patch Dataset				
BIRADS	Normal	Benign	Malignant	Total
d	12	14	11	**37**
e	28	1	5	**34**
f	24	8	6	**38**
g	26	9	6	**41**
Total	**90**	**32**	**28**	**150**
IRMA: DDSM Patch Dataset				
d	203	219	222	**644**
e	168	232	228	**628**
f	195	225	227	**647**
g	207	224	226	**657**
Total	**773**	**900**	**903**	**2576**

Some of these related works are summarized in Table 5.2, along with the details of feature vectors, classifiers, and a number of images used in experiments. The accuracy

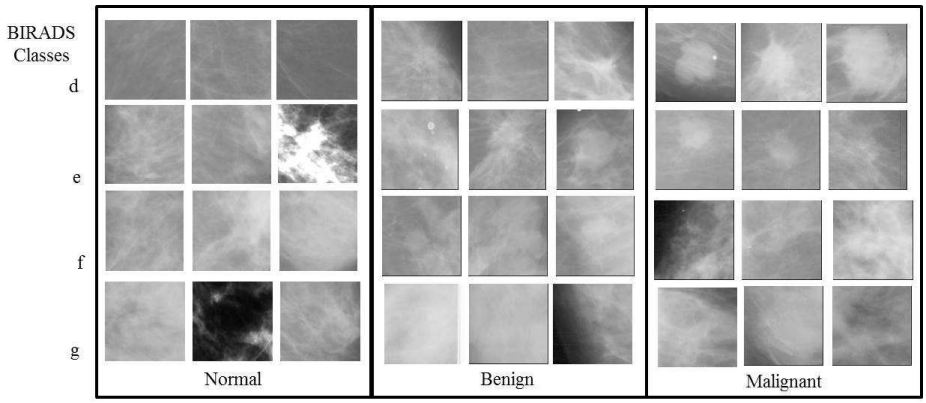


Figure 5.1: Samples of mammogram patches from the IRMA database. Row denotes the density of patches.

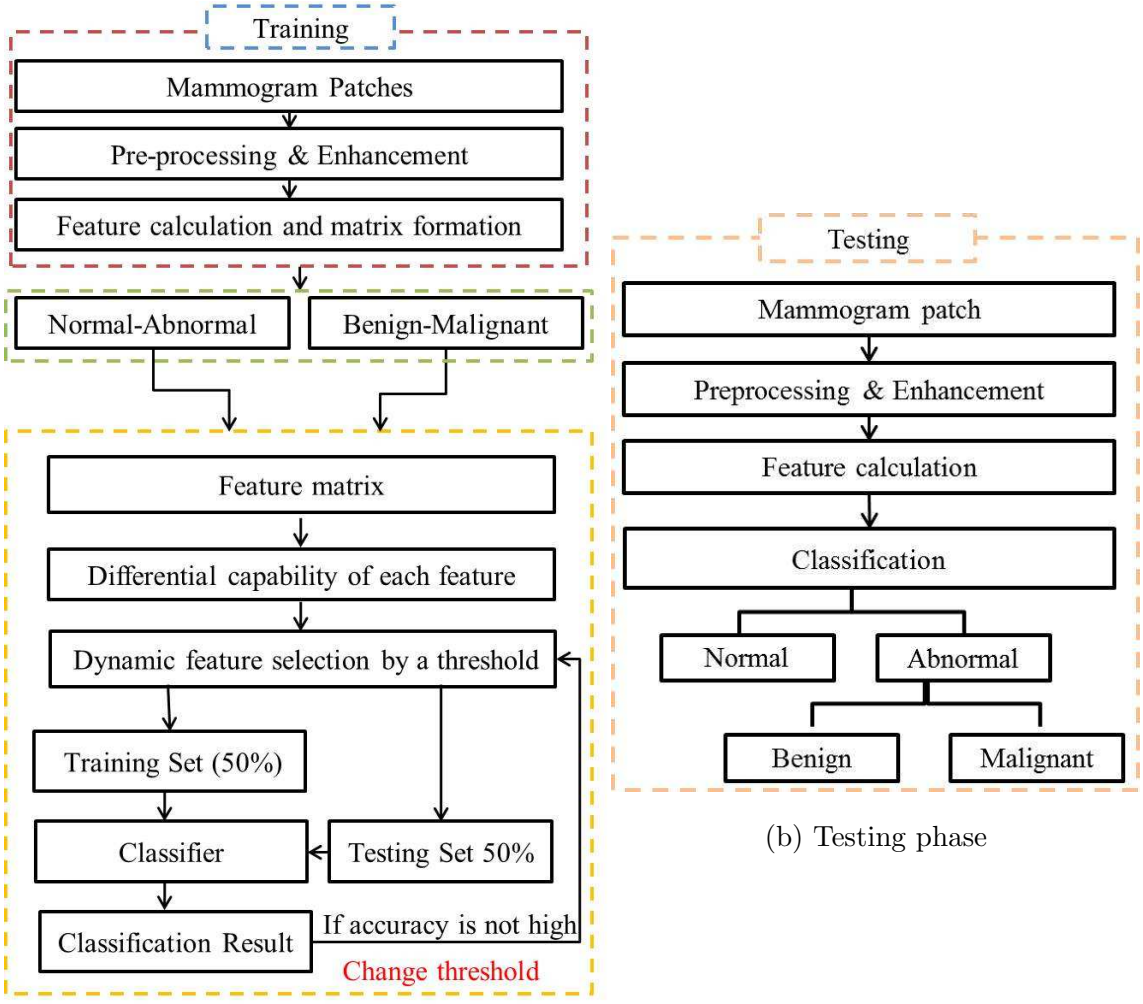
obtained for both types of classification (normal–abnormal and benign–malignant) is also given. The performance of approaches depends on different factors of mammogram patches such as dimension, number of training and testing samples, resolution and the type of abnormality. Most of the works in this area have tested on a subset of images instead of all mammogram patches, which we use.

Table 5.2: Summary of some related works on mammogram patch classification.

Feature	**Classifier**	**Database**	**# of images**	**Accuracy (normal-abnormal)**	**Accuracy (benign-malignant)**
Gabor + PCA [85]	SVM	DDSM	NA	84.00%	78.00%
GLCM + DWT [80]	BPNN	MIAS	332	98.10%	95.04%
		DDSM	550	99.45%	97.61%
HOG, DSIFT, & LCP [31]	SVM	DDSM	600	84.00%	78.00%
FFST [86]	SVM	MIAS	228	98.29%	100.00%
		DDSM	228	100.00%	98.29%
Gabor [96]	PSO + SVM	DDSM	1024	98.82%	91.61%

5.2 Proposed Mammogram Patch Classification System

As mentioned earlier, this work proposes a two-stage mammogram patch classification system. In the first stage, mammogram patches are classified as normal–abnormal, and in the second stage, abnormal mammogram patches are further classified as benign–malignant. The framework of the proposed work for training and testing phase is shown in Figure 5.2.



(a) Training phase

(b) Testing phase

Figure 5.2: Flow diagram of the proposed mammogram patch classification system.

Here, we first discuss image pre-processing and enhancement techniques used by

us. Mammogram patches are preprocessed for illumination normalization and visibility enhancement of tumors and tissues [23, 25]. A two-stage adaptive histogram equalization enhancement technique is used here for texture enhancement of mammogram patches [23]. Second, we discuss our two proposed feature extraction techniques, where features of mammogram patches are extracted from enhanced images. Finally, we discuss the feature selection technique used by us.

5.2.1 Pre-processing and Enhancement

In this work, for pre-processing, we only normalize the intensity of pixels and that too for a few mammogram patches. This is because while capturing images, illumination conditions are usually not the same. So, the range of the gray level is different for different mammogram patches. Hence, we use a simple and the most commonly used normalization formula (given below), which normalizes the intensity of pixels between 0 and 1 [72, 73]

$$I'(x,y) = \frac{I(x,y) - \min(I)}{\max(I) - \min(I)}.$$

where (x, y) is the pixel position, $I'(x, y)$ is the normalized pixel intensity, $I(x, y)$ is the actual pixel intensity, $\min(I)$ is the minimum intensity over all the pixels, and $\max(I)$ is the maximum intensity over all the pixels.

Next, we discuss tissue enhancement of mammogram patches. Histogram equalization is the one of the most basic technique here, which stretches the contrast of the high histogram regions and compresses the contrast of the low histogram regions. As a result, if the region of interest in an image occupies only a small portion, it will not be properly enhanced during histogram equalization.

This leads to more advanced techniques for enhancement, e.g., Adaptive Histogram Equalization (AHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), Unsharp Masking (UM), Non-Linear Unsharp Masking (NLUM), Two-Stage Adaptive Histogram Equalization (TSAHE), etc. [23, 102].

CLAHE has been found to be more suitable for tissue enhancement in mammogram patches. One aspect of enhancement is to capture the texture of the breast in

mammogram patches better, which is defined by entropy. In the work of [23], authors have shown that TSAHE performs better than most of the existing techniques in not just the overall enhancement (defined by EME, i.e. Measure of Enhancement) but entropy as well.

Since cancerous cells mostly develop in tissues and our proposed descriptors (HOT and PB-DCT) are strongly tied to breast texture, we use a combination of CLAHE and TSAHE. We term it as TS-CLAHE.

We apply two stages of CLAHE on mammogram patches in a cascaded order. Firstly, histogram equalization is applied to 8×8 sized blocks, followed by an application to 4×4 sized of blocks. Fig. 5.3 shows the normalized and the enhanced image of a mammogram patch. It is observed that mass tissues are clearly visible in the enhanced image.

Figure 5.3: A preprocessed and enhanced mammogram patch.

5.2.2 Feature Extraction Techniques

As discussed in the previous sections, we propose two descriptors (HOT and PB-DCT) for mammogram patch classification. HOT is a modification of the HOG descriptor where a Gabor filter is used to calculate the angle and the magnitude response of texture of a mammogram patch. Selected PB-DCT coefficients based features are used here to improve the classification accuracy for each density class. Next, we discuss these two techniques separately. To the best of our knowledge, these strategies have not been applied anywhere for mammogram patch classification.

5.2.2.1 The Histogram of Oriented Texture

Here, we derive our HOT descriptor. Firstly, we discuss the calculations of gradient and orientation of an image as well as the HOG descriptor calculation from cells and blocks partitions [31]. Secondly, we describe a Gabor filter, which is used to extract magnitude and orientation of tissue texture information, and finally, we discuss modifications to the HOG descriptor that involves a Gabor filter and parameter selection.

Gradient of an image I in horizontal and vertical directions, for a pixel position (x, y) is computed as

$$\begin{aligned} dx &= I(x+1, y) - I(x-1, y) \text{ and} \\ dy &= I(x, y+1) - I(x, y-1), \end{aligned}$$

respectively. For each pixel, $I(x, y)$, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are computed as below.

$$m(x, y) = \sqrt{dx^2 + dy^2} \text{ and} \tag{5.1}$$

$$\theta(x, y) = tan^{-1}\left(\frac{dy}{dx}\right). \tag{5.2}$$

Orientation range $(0^{\circ} - 180^{\circ})$ is quantized into B bins (i.e. $\theta(x, y) \in bin(b)$ with $b = 1, 2, 3, \ldots, B$). The image is divided into $c \times c$ non-overlapping cells, and $l \times l$ cells are integrated as one block. Two adjacent blocks can overlap. The histogram of orientations $(HC(b)_i)$ of bin(b) within i^{th} cell is computed as

$$\begin{aligned} HC(b)_i &= HC(b)_i + m(x, y), \\ m(x, y) &\in Cell_i, \\ b &= 1, 2, 3, \ldots, B, \text{ and} \\ i &= 1, 2, 3, \ldots, c \times c. \end{aligned}$$

The histogram of j^{th} block (HB_j) is obtained by integrating HCs (Histogram of Cells) within this block as follows:

$$HB_j = HC_1 \| HC_2 \| \ldots \| HC_{l \times l},$$

where $\|$ denotes histograms concatenation into a vector. The vector of HB_j is finally normalized by L_2-norm block normalization as below to obtain NHB_j.

$$NHB_j = \frac{HB_j}{\sqrt{\|HB_j\|_2^2 + e^2}},$$

where e is a small constant to avoid problem of division by zero. HOG can be obtained by integrating normalized histograms of all blocks as below.

$$HOG = NHB_1 \| NHB_2 \| \dots NHB_j \| \dots \| NHB_N,$$

where N is the number of possible blocks in an image, which is equal to $(c - l + 1) \times (c - l + 1)$. Fig. 5.4 shows an example of cell partitions, formation of overlapped blocks and concatenation of histograms to get the HOG descriptor. Finally, the length of HOG is $l^2 \times (c - l + 1)^2 \times B$.

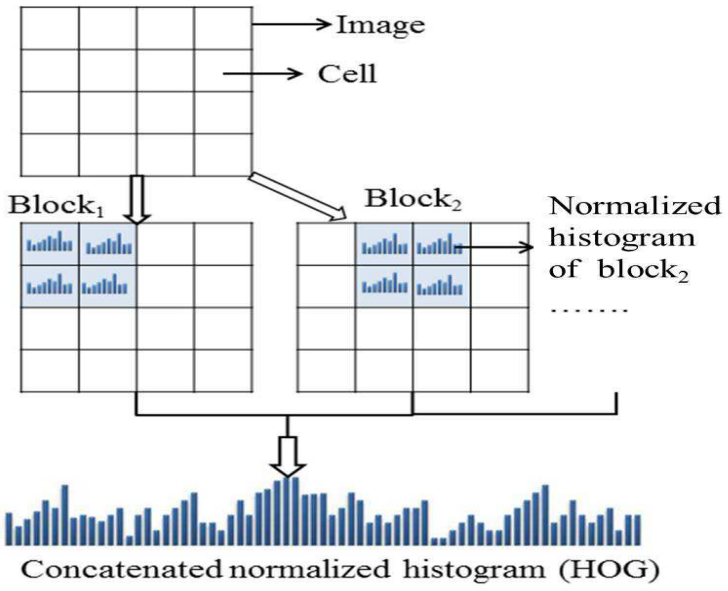


Figure 5.4: The HOG descriptor calculation.

Different line-shape filters or tools are available in the literature to extract lines and orientation features of a texture image [96]. 2-D Gabor filters have been found more suitable filter bank to extract biological-like textural features of simple cells in the mammalian visual system [103]. Thus, a Gabor filter is ideal for calculating multi-orientation texture features of a mammogram patch. A Gabor function is defined as

follows:

$$G(x, y, \theta, \mu, \sigma) \quad = \quad \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} \exp\left[2\pi i \left(\mu x \cos\theta + \mu y \sin\theta\right)\right],$$

where $i = \sqrt{-1}$, μ is the frequency of the sinusoidal wave, θ controls the orientation of the function, and σ is the standard deviation of the Gaussian envelop. Based upon this Gabor function, a set of Gabor filters can be created for different scales and orientations.

Here, texture feature extraction for a given mammogram patch image (I) is calculated by the real part of a Gabor filter bank with eight different orientations and a fixed scale. Gabor magnitude, $m(x, y)_{Gabor}$, and Gabor orientation, $\theta(x, y)_{Gabor}$, of each pixel (x, y) are computed as

$$m_{Gabor}(x, y) = min(I(x, y) * G(x, y, \theta_t, \mu, \sigma)) \quad \text{and}$$

$$\theta_{Gabor}(x, y) = argmin_t(I(x, y) * G(x, y, \theta_t, \mu, \sigma)),$$

where $*$ means the convolution operation. The direction θ_t is calculated as follows:

$$\theta_t = \frac{\pi(t-1)}{8}, \; t = 1, 2, \ldots, 8.$$

The features are calculated by varying the values of σ and μ. Fig. 5.5 shows the magnitude and angle image of a mammogram patch.

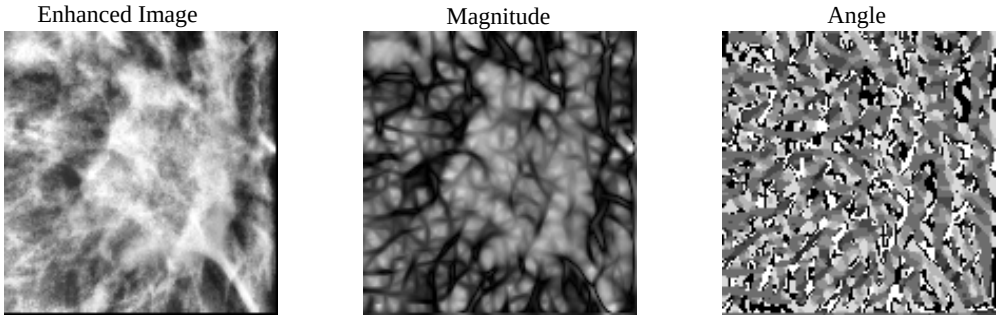


Figure 5.5: Gabor magnitude and angle image.

We combine HOG with a Gabor filter and name it as HOT. HOT is computed in the same way as HOG, but $m_{Gabor}(x, y)$ and $\theta_{Gabor}(x, y)$ are used as magnitude and orientation of texture line instead of (5.1) and (5.2), respectively.

Finally, optimum parameters of the HOT descriptor, for both types of classification, are chosen by experiments. The value of σ is varied from one to five to obtain an optimum number. The value of μ is computed as $\frac{1}{\sqrt{2\sigma}}$. Here, the magnitude image is divided into equal sized 16×16 cells. Size of a block considered is 2×2, therefore, 15×15 overlapped blocks are formed. The orientation range $(0^{\circ} - 180^{\circ})$ is quantized into 8 bins, and therefore, the final length of the resultant HOT descriptor is 7200. The length of the HOT descriptor is large, and all features do not have same discrimination capability [104]. Feature selection schemes help to select more appropriate features. This is discussed in Section 5.2.3.

Next, we compare our proposed HOT descriptor with other works that use a combination of HOG and Gabor filter. In work by [99], authors use a Gabor filter to extract features and HOG to reduce the dimension of the extracted feature vector. However, we use a combination of both to extract features. Comparison with two other works in this area is given in Table 5.3. Apart from the differences discussed until now, we (i.e. in the HOT descriptor) use DP for feature selection, which none of the other works use.

Table 5.3: Comparison of various descriptors using a combination of HOG and Gabor filter

Parameters	**HoGG**	**Gabor-HOG**	**HOT**
# of Orientations	9	4	8
# of Bins	9	9	8
# of Cells to Divide Image	4×8	Retangular Cells (Number not given)	16×16
Overlapped Image Area	3×7	Half of the Image	15×15
Feature Vector Length	756	81	7200
Application Area	Human Detection	Face Recognition	Mammogram Classification

5.2.2.2 Pass Band - Discrete Cosine Transform

2D Discrete Cosine Transform (DCT) transforms images into frequency representation from the spatial form. It also provides energy compaction, that helps to reduce the information redundancy by retaining only a few coefficients. DCT coefficients of $I(x, y)$ image of $M \times N$ size are calculated as follows:

$$F(u,v) = \frac{1}{\sqrt{MN}}\alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y) \times \cos\left(\frac{(2x+1)u\pi}{2M}\right) \times cos\left(\frac{(2y+1)v\pi}{2N}\right)$$
$$\text{with } u = 0, 1, 2, \ldots, M,\ v = 0, 1, 2, \ldots, N,$$

where $\alpha(\omega)$ is defined by

$$\alpha(\omega) = \begin{cases} \frac{1}{\sqrt{2}} & \omega = 0 \\ 1 & \text{otherwise.} \end{cases}$$

DCT based various feature extraction and compression techniques have been proposed in literature [88]. Usually, DCT features are formed by selecting the most prominent and discriminating coefficients based upon some criterion [88]. The DCT coefficients can be divided into three sets, low frequencies, middle frequencies and high frequencies. Low frequencies are correlated with illumination conditions, middle frequencies represent texture features, while high frequencies represent small variance or noise. Illumination and texture properties are important for mammogram patch classification. Therefore, this work uses low and middle coefficients to form the descriptor (the Pass Band - Discrete Cosine Transform descriptor or abbreviated as PB-DCT). Finally, the more discriminate DCT coefficients are selected based upon a discrimination criterion, which is discussed in the next section.

5.2.3 Feature Selection with Discrimination Potentiality

All features do not have the same ability to discriminate various classes (normal-abnormal and benign-malignant) [104], and they do not increase the accuracy based on available information for each class. Therefore, it is necessary to eliminate irrelevant features and select the most discriminative features among a given set of features [105].

Determining features to improve accuracy as well as reduce searching time is a difficult task. As discussed in Section 5.1, feature subset selection techniques have been found to be more suitable for mammogram patch classification as compared to dimension reduction techniques.

There exist many techniques for feature selection. Some of the common ones are as follows: PCA (Principal Component Analysis) method, Markov blanket method, wrapper methods (e.g., sequential selection algorithm, genetic algorithms etc.), filter methods (e.g., Pearson correlation criteria, mutual information etc.), embedded methods, and statistical measures based methods (e.g., T-test, Kolmogorov-Smirnov test, Kullback-Leibler divergence etc.) [106, 107].

Out of these, wrapper methods, filter methods, and statistical measures based methods are usually used for mammogram patch classification. Wrapper methods are computationally expensive since the number of steps required for obtaining the feature subset are very high. Filter methods sometimes lead to a redundant feature subset, and hence, are not optimal in this sense. [106]. Thus, we go for statistical measures based methods since they do not have the above discussed drawbacks. These methods also have the advantage in reducing the feature space without significantly degrading the classification performance [108].

The T-test method is one such method that gives a high score to features that capture the texture and the shape of mammograms [109]. As earlier, capturing texture is very important to us. Moreover, the T-test method is computationally light, easy to implement, and has been very recently successfully applied in mammogram context [86, 24]. Thus, we use this feature selection method and show in the results section that this works very well with our proposed descriptors (DP-HOT and DP-PB-DCT). We term it as DP because of its capability in discriminating between the available features.

The discrimination potentiality DP_k of the k^{th} feature between two classes (a and b) is computed from a given training set as follows:

$$DP_k = \frac{\mu_{a,k} - \mu_{b,k}}{\sqrt{\frac{\delta_{a,k}^2}{n_a} - \frac{\delta_{b,k}^2}{n_b}}},$$

where $\mu_{a,k}$, $\mu_{b,k}$, and $\delta_{a,k}$, $\delta_{b,k}$, are mean and standard deviation values of the k^{th} feature for a and b classes, respectively. n_a and n_b are the number of mammogram patches for a and b classes, respectively. A high value of DP means high discrimination ability of the corresponding feature [24].

All features (columns) of the feature matrix are arranged in descending order of their DP value. Initially, first five features with highest DP values are chosen for classification accuracy. Then, classification accuracy is calculated by adding features, with next higher value of DP, one by one until we get the highest accuracy. The optimum subset of features corresponding to the highest accuracy is selected as the final descriptor.

5.3 Experimental Results

Experiments are carried out in MATLAB® 2016 on a machine with Intel i5 processor @ 2.5 GHz and 4GB RAM. As discussed earlier, mammogram patches are taken from the IRMA database (the MIAS and DDSM datasets). All images from this database (density wise), as given earlier in Table 5.1, are used. The size of each mammogram patch is 128×128.

We *first* use two-fold cross-validation, where the dataset is randomly divided into two equal parts. One part is used for training and the other is used for testing. Then, the two parts are swapped. That is, the one used for training earlier is now used for testing, and the one used for testing earlier is now used for training. At the end of this exercise, average performance is saved. *Finally*, we repeat two-fold cross-validation ten times so as to remove any bias related to the division of the dataset. Use of two-fold cross-validation and repeating it ten times ensures that the classification system is not over-fitted.

The performance of our system (and comparative systems) is evaluated by standard metrics of sensitivity, specificity, accuracy and AUC (Area Under the ROC Curve).

Sensitivity is computed as the number of true positive cases over the number of

actual positive cases. It is represented as follows:

$$Sensitivity = \frac{TP}{TP + FN}(\%),$$

where TP means True Positive cases and FN means False Negative cases.

Specificity is computed as the number of true negative cases over the number of actual negative cases. It is represented as follows:

$$Specificity = \frac{TN}{FP + TN}(\%),$$

where TN means True Negative cases and FP means False Positive cases.

Accuracy is computed as the number of correct classifications over the number of given cases. It is represented as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}(\%).$$

AUC (Area Under the ROC Curve) provides a measure of the overall performance of the classifier, i.e. larger the area, better the classification. It is calculated from the trapezoidal rule as below [110].

$$AUC = \frac{1}{2}\sum_{k=1}^{n}((Spec(k) - Spec(k+1)) * (Sens(k) + Sens(k+1)))(\%),$$

where n is the total number of test cases, and $Spec(k)$ and $Sens(k)$ are Specificity and Sensitivity for the k^{th} test case, respectively. In the case of an ideal classification system, the value of all the four metrics should be close to 100%.

This section has three subparts. In Section 5.3.1, experiments related to finding the optimum parameters of the HOT descriptor with DP, from now on referred as DP-HOT, are given. Similar experimental results for the PB-DCT descriptor with DP, from now on referred as DP-PB-DCT, are given in Section 5.3.2. Finally, in Section 5.3.3, the performance of both our proposed descriptors, DP-HOT and DP-PB-DCT, is compared with the performance of the existing (and popular) descriptors.

For fixing the parameters, only the training data is used (i.e. 50%, as we use two-fold cross-validation) and the validation data for testing is kept blind.

5.3.1 Performance of DP-HOT

Firstly, experiments are performed to find optimum parameters of Gabor filter for all the classes (i.e. d, e, f, g, and "all")[1]. Performance parameters are calculated by varying the value of σ from one to five to obtain suitable scale for each density class. The best accuracy obtained by varying σ is mentioned here.

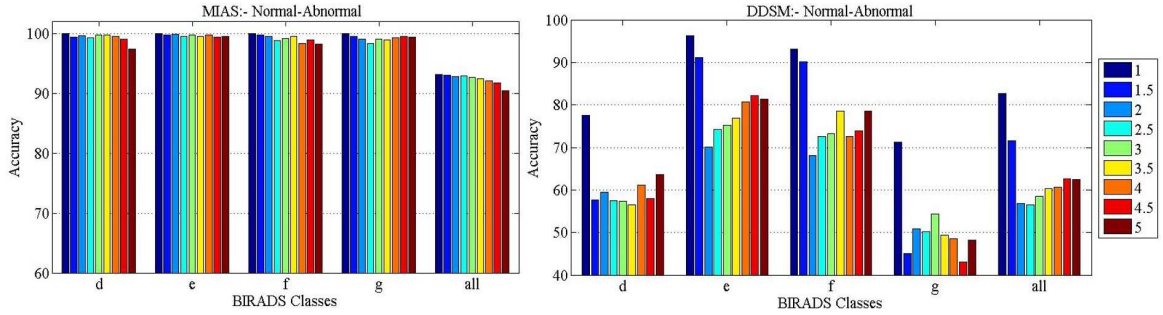


Figure 5.6: Comparison of normal-abnormal classification accuracies obtained by varying the value of σ from one to five for each individual BIRADS class.

Fig. 5.6 compares normal-abnormal classification accuracy for different values of σ for each class. It is observed that the DP-HOT descriptor achieves approximately 100% accuracy for all values of σ for all the classes of the MIAS dataset. The maximum normal-abnormal classification accuracy of all the classes combined is achieved with σ as one. It is difficult to infer the optimum value of σ by observing the bar chart of the MIAS dataset only. In case of the DDSM dataset, the DP-HOT descriptor with σ as one gives the best accuracy for all the classes individually as well as combined. For classes e and f, DP-HOT achieves an accuracy of around 95%, while it does not achieve good accuracy for classes d and g (around 70%).

Fig. 5.7 compares benign-malignant classification accuracy for different values of σ for each class. The DP-HOT descriptor again achieves approximately 100% accuracy for all values of σ for all the classes of the MIAS dataset. The maximum benign-malignant classification accuracy for all the classes combined is achieved with σ as 3. For the DDSM dataset, the DP-HOT descriptor performs equally for all σ for all the

[1]For the "all" class, we have combined all the images from d, e, f, and g class and this is available from the dataset itself.

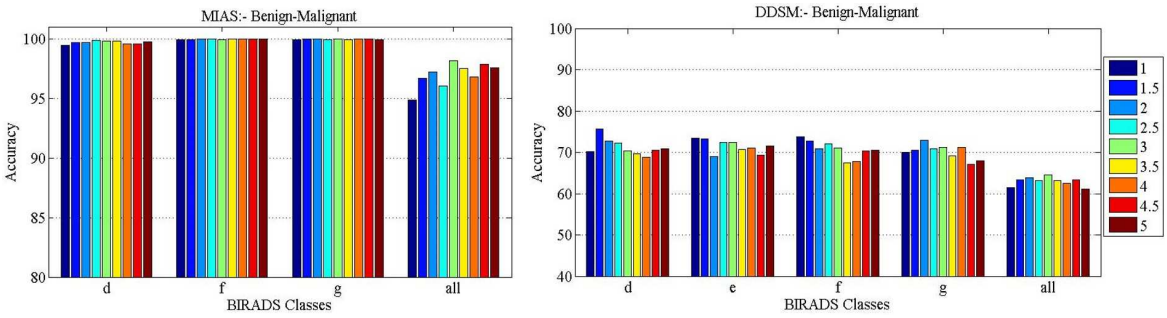


Figure 5.7: Comparison of benign-malignant classification accuracies obtained by varying the value of σ from one to five for each individual BIRADS class.

classes (around 70%).

First observation is that DP-HOT does not perform very well for both types of classification (normal-abnormal and benign-malignant) for the DDSM dataset (although it does perform well for MIAS). The reason, as discussed in Section 5.1, is that texture information plays a big role in both types of classifications (normal-abnormal and benign-malignant), and DP-HOT does not capture that well for the DDSM dataset. The DP-PB-DCT descriptor overcomes this drawback to a great extent. Another observation is that the classification accuracy for an individual class is better as compared to the accuracy for the "all" class. This is intuitive because images from an individual class have similar features (they belong to same density class). On the other hand, images from the "all" class have features that vary substantially (again, due to varying density). For all the forthcoming experiments, a similar behavior is observed for both the descriptors.

Table 5.4 lists the feature length used for both types of classification (normal-abnormal and benign-malignant) done on both the datasets (MIAS and DDSM) for each density class.

5.3.2 Performance of DP-PB-DCT

As in the case of the DP-HOT descriptor, in Table 5.5 we list the feature length used for both types of classification (normal-abnormal and benign-malignant) done on both the datasets (MIAS and DDSM) for each density class when using the DP-PB-

Table 5.4: Feature length for the DP-HOT descriptor.

BIRADS Class	MIAS: Feature Length	DDSM: Feature Length
Normal-Abnormal		
d	2655	210
e	10	100
f	30	75
g	20	200
All	40	140
Benign-Malignant		
d	2000	225
e	All Benign	130
f	2000	405
g	1790	170
All	2000	990

DCT descriptor.

Since, the selection of features in the DP-PB-DCT descriptor is more critical, we do further feature selection analysis here. Fig. 5.8 compares accuracy against the number of features for normal-abnormal and benign-malignant classification for each density class separately and combined. As in the case of DP-HOT, the performance for individual density is better than combined. Moreover, multiple points with the high classification accuracy are observed.

5.3.3 Comparison with Other Techniques

The performance of the two proposed descriptors is compared with some related descriptors such as Zernike moment [27], Multiple LPQ (MLPQ) [28], GRsca [29], Wavelet Gray Level Co-occurrence Matrix (WGLCM) [30], LCP [31] and HOG [31] for each density class separately as well as combined. The performance of each descriptor is evaluated in the same experimental setup including the use of the same testing protocol.

Table 5.5: Feature length for the DP-PB-DCT descriptor.

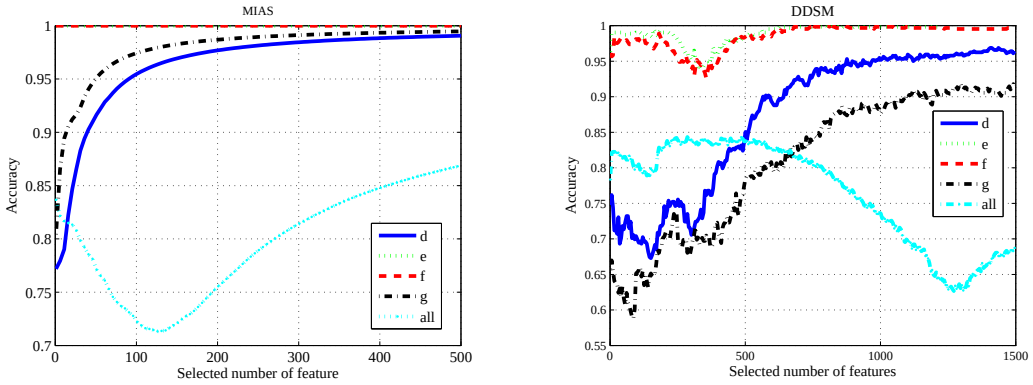
BIRADS Class	**MIAS: Feature Length**	**DDSM: Feature Length**
Normal-Abnormal		
d	5	4000
e	5	4000
f	5	995
g	5	1315
All	5	430
Benign-Malignant		
d	5	4350
e	5	4005
f	5	3995
g	5	4625
All	5	5065

The parameters of each of these descriptors are selected as given in the literature to achieve the best performance for mammogram patch classification. These parameters are summarized below.

The Zernike moment descriptor is computed by dividing a mammogram patch into 4×4 blocks. 120 Zernike moments are used for obtaining the final mammogram patch descriptor. Therefore, the length of the feature vector is $4 \times 4 \times 120$ [27].

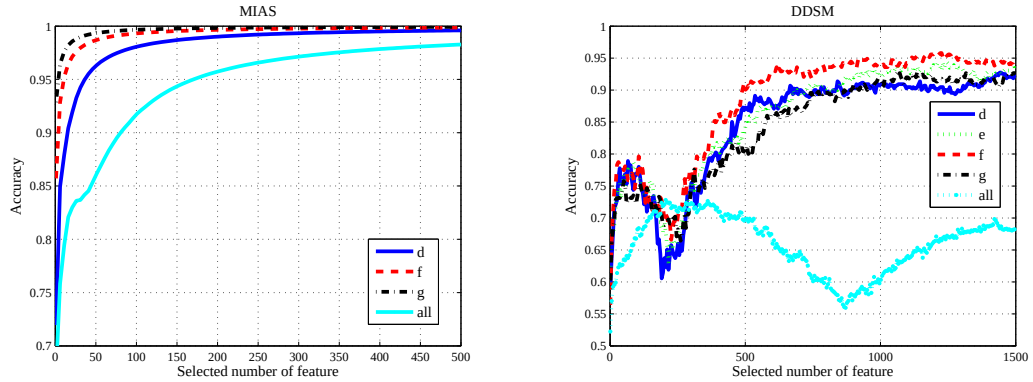
The MLPQ descriptor is computed by concatenating different LPQ descriptors. Each LPQ descriptor is obtained by varying values of three different parameters. This includes filter size (r with values 1, 3, 5), the scalar frequency (a with values 0.8, 1.0, 1.2, 1.4, 1.6), and the correlation coefficient (ρ with values 0.75, 0.95, 1.15, 1.35, 1.55, 1.75, 1.95). Each LPQ descriptor's length is standard (256). Therefore, the length of the feature vector is $3 \times 5 \times 7 \times 256$ [28].

The GRsca descriptor extracts features from the three gray-level run length co-occurrence matrices corresponding to the original image and two filtered images (using a filter of size 3 and 5). These images are divided into four blocks. Features are ob-



(MIAS) (DDSM)

Normal-Abnormal



(MIAS) (DDSM)

Benign-Malignant

Figure 5.8: Performance accuracy against the number of DP-PB-DCT features.

tained from co-occurrence matrix and these four blocks. A set of ten different descriptors is calculated at four different orientations and two different distances. Therefore, the length of the feature vector is $3 \times 5 \times 10 \times 4 \times 2$ [29].

The WGLCM descriptor is computed by decomposing the image into one approximation. Detail coefficients up to two levels with a wavelet filter are used. For each level, three decompositions (along horizontal, vertical and diagonal directions) are obtained. The normalized GLCM (NGLCM) matrices of all detail coefficients are calculated in four directions (0°, 45°, 90°, and 135°) with a single displacement. Contrast, homogeneity, energy, and correlation statistical properties of all NGLCM matrices are

calculated and concatenated into a vector. Therefore, the length of the feature vector is $2 \times 3 \times 4 \times 4$ [30].

The LCP is a modification of Local Binary Pattern (LBP). Here, first, the weights associated with intensities of neighboring pixels are used to linearly reconstruct the central pixel intensity. Then, the error between the central pixel and its neighbor is minimized. In this work, LCP images are computed for radius 1 to 5. Each LCP image is divided into 4×4 blocks. The histogram of each block is concatenated with 58 bins, and the result is used as a mammogram patch descriptor. Therefore, the length of the feature vector is $5 \times 4 \times 4 \times 58$ [31].

The HOG is calculated using 16×16 cell partitions. The size of the block considered is 2×2, thus, 15×15 overlapped blocks are formed. The orientation range is quantized into 8 bins. Therefore, the length of the feature vector is $2 \times 2 \times 15 \times 15 \times 8$ [31].

The optimum parameters of DP-HOT and DP-PB-DCT are selected based upon experiments as discussed in the previous subsections. The length of each descriptor is different and the appropriate feature set for each descriptor is selected based upon the DP values as described earlier.

Table 5.6 compares sensitivity, specificity, accuracy and AUC for normal-abnormal classification with all the above descriptors on the MIAS and DDSM datasets. The performance parameters for all the descriptors are provided for each density class separately as well as combined. The best performing systems are highlighted in bold.

For the MIAS dataset, both the proposed descriptors (DP-HOT and DP-PB-HOT) achieve near 100% sensitivity, specificity, accuracy, and AUC for all the classes. This is better than the six standard descriptors (Zernike moment, MLPQ, GRsca, WGLCM, LCP, and HOG).

For the DDSM dataset, although our DP-HOT descriptor performs almost as badly as the other six descriptors for all the classes (as low as around 65% for one performance parameter), DP-PB-DCT performs extremely well for all the classes (more than 92% for most performance parameters), which is better than the six standard descriptors. All the eight descriptors (the six standard and the two new) perform well for classes e and f but have a dip in the performance for classes d and g. This is because for the

Table 5.6: Mammogram patch classification results as normal-abnormal for the MIAS and DDSM datasets.

Descriptor	**BIRADS**	**MIAS**				**DDSM**			
		Sens.	**Spec.**	**Acc.**	**AUC**	**Sens.**	**Spec.**	**Acc.**	**AUC**
Zernike [27]	*d*	99.12	93.59	97.32	90.28	90.01	68.64	83.32	91.69
	e	94.98	99.63	98.81	**100**	99.78	85.12	95.86	**100**
	f	92.86	99.88	97.29	**100**	99.56	95.91	98.46	**100**
	g	91.31	97.87	95.42	75.82	83.32	36.72	68.63	81.89
	All	75.92	91.70	85.39	85.11	**91.90**	61.99	82.95	82.03
MLPQ [28]	*d*	89.21	75.00	84.70	90.28	80.29	73.51	78.16	89.71
	e	**100**	96.30	96.96	**100**	99.97	99.97	99.97	**100**
	f	**100**	**100**	**100**	**100**	99.91	99.91	99.91	**100**
	g	68.75	92.31	83.10	90.66	83.14	54.06	73.97	81.80
	All	72.60	88.50	82.14	88.15	86.07	88.14	86.69	93.59
GRsca [29]	*d*	76.92	70.00	74.59	85.90	86.03	79.53	84.00	88.23
	e	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	f	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	g	66.96	**100**	87.86	84.13	83.41	63.85	77.24	82.26
	All	68.95	98.79	86.85	91.26	88.44	88.64	88.50	**95.48**
WGLCM [30]	*d*	92.15	50.83	78.71	73.61	85.54	84.92	85.34	94.38
	e	90.91	**100**	98.40	**100**	99.84	99.98	99.87	**100**
	f	**100**	**100**	**100**	**100**	99.70	**100**	99.79	**100**
	g	60.96	**100**	85.93	80.77	84.94	62.25	77.78	85.81
	All	64.72	**100**	85.89	93.41	86.24	93.21	88.33	**94.97**
LCP [31]	*d*	92.31	66.67	83.63	80.56	87.99	89.07	88.32	90.39
	e	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	f	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	g	73.21	96.15	87.74	64.42	86	69.10	80.67	74.93
	All	75	97.78	88.67	85.70	87.80	**93.39**	**89.47**	92.88
HOG [31]	*d*	88.46	91.67	89.47	**100**	84.36	78.62	82.56	87.58
	e	**100**	**100**	**100**	**100**	99.57	94.64	98.25	98.79
	f	**100**	95.83	97.36	**100**	98.23	88.73	95.36	97.67
	g	85.71	**100**	95	**100**	82.22	56.52	74.12	77.70
	All	70	87.78	80.67	85.85	87.69	77.95	84.77	88.32
DP-HOT	*d*	**100**	**100**	**100**	**100**	83.44	64.69	77.57	91.20
	e	**100**	**100**	**100**	**100**	96.96	95.24	96.50	**100**
	f	**100**	**100**	**100**	**100**	94.69	89.75	93.20	**100**
	g	**100**	**100**	**100**	**100**	77.78	57.06	71.24	76.05
	All	87	98	93	97.26	84.80	77.30	82.56	86.21
DP-PB-DCT	*d*	**100**	**100**	**100**	**100**	**98.19**	**94.01**	**96.88**	**98.39**
	e	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	f	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	g	**100**	**100**	**100**	**100**	**97.78**	**80.70**	**92.39**	**98.68**
	All	**97.18**	98.89	**97.33**	**100**	87.18	79.37	84.84	92.20

DDSM dataset, texture discrimination for e and f classes is better than for d and g classes, which is crucial in normal-abnormal classification.

Table 5.7 compares performance parameters for benign-malignant classification with all the above eight descriptors for all the classes (and combined) of the MIAS and DDSM datasets. As earlier, the best performing systems are highlighted in bold. The results are similar to those for normal-abnormal classification. Note that in the MIAS dataset, the e class has all benign images, so, classification was not performed for this. The corresponding rows in this table are left empty.

For the MIAS dataset, both the proposed descriptors (DP-HOT and DP-PB-DCT) perform slightly better than the six standard descriptors for all the classes (achieve near 100% sensitivity, specificity, accuracy, and AUC).

For the DDSM dataset, DP-HOT is slightly better than the existing descriptors (around 70% for all performance parameters), while DP-PB-DCT is much better than the six standard descriptors for all the classes (more than 92% for most performance parameters).

To summarize, as mentioned in Section 5.1 as well as Section 5.3.1, capturing texture information is of utmost importance for mammogram patch classification (both normal-abnormal and benign-malignant). In general, DP-HOT captures this texture information slightly better than the six standard descriptors, and hence, it performs slightly better than these. DP-PB-DCT captures texture information best, and hence, performs much better than all the others.

Table 5.7: Mammogram patch classification results as benign-malignant for the MIAS and DDSM datasets.

Descriptor	**BIRADS**	**MIAS**				**DDSM**			
		Sens.	**Spec.**	**Acc.**	**AUC**	**Sens.**	**Spec.**	**Acc.**	**AUC**
Zernike [27]	*d*	93.23	97.05	95.39	54.76	60.36	54.34	57.37	56.78
	e	-	-	-	-	60.09	60.34	60.22	58.39
	f	97.26	99.83	98.73	**100**	58.62	56.02	57.30	63.26
	g	95.61	98.96	97.62	**100**	57.96	70.98	64.44	57.22
	All	80.47	84.15	82.44	72.77	56.92	55.11	56.02	53.87
MLPQ [28]	*d*	71.43	75.00	71.79	96.43	59.59	57.43	58.50	62.70
	e	-	-	-	-	61.72	59.36	60.55	63.24
	f	96.43	97.62	96.94	**100**	49.79	64.76	57.31	59.18
	g	100	83.33	93.75	93.34	**95.98**	17.70	56.67	57.66
	All	73.96	76.19	75.00	86.16	57.40	54.62	56.01	56.04
GRsca [29]	*d*	64.29	58.33	60.26	62.86	66.83	54.21	60.47	62.01
	e	-	-	-	-	71.09	53.94	52.59	66.24
	f	85.71	28.57	61.22	58.34	63.59	56.54	60.05	64.32
	g	77.59	58.28	69.49	50	81.70	40.27	60.89	64.58
	All	96.88	35.71	68.33	68.75	67.69	52.63	60.15	64.21
WGLCM [30]	*d*	35.83	92.86	67.95	80.95	53.15	62.99	58.04	60.35
	e	-	-	-	-	50.88	63.79	57.39	56.32
	f	45.83	90.63	71.43	91.67	59.92	52.01	55.97	59.24
	g	47.22	86.25	70.83	93.34	57.08	60.71	58.89	57.59
	All	53.81	74.72	64.96	70.09	53.20	56.98	55.09	56.59
LCP [31]	*d*	99.72	99.84	99.77	**100**	51.80	66.66	59.18	63.38
	e	-	-	-	-	57.89	62.93	60.43	60.86
	f	**100**	99.9	99.98	**100**	51.96	62.24	57.08	60.33
	g	99.99	99.96	99.98	**100**	56.64	61.16	58.89	60.04
	All	98.94	95.57	97.37	72.32	55.26	59.33	57.29	57.38
HOG [31]	*d*	**100**	**100**	**100**	95.24	64.86	65.30	65.08	66.27
	e	-	-	-	-	59.65	65.52	62.61	70.01
	f	**100**	**100**	**100**	**100**	66.95	65.32	66.14	67.92
	g	**100**	**100**	**100**	**100**	63.27	66.07	64.67	68.07
	All	80.67	92.86	87.5	94.20	54.93	61.89	58.40	65.06
DP-HOT	*d*	**100**	**100**	**100**	**100**	72.07	68.49	70.29	80.44
	e	-	-	-	-	70.18	75	72.61	74.92
	f	**100**	**100**	**100**	**100**	72.66	69.31	71.02	83.91
	g	**100**	**100**	**100**	**100**	73.45	68.75	71.11	81.09
	All	98.83	97.57	98.24	**100**	64.56	64.67	64.61	68.89
DP-PB-DCT	*d*	**100**	**100**	**100**	**100**	**96.85**	**94.53**	**95.69**	**98.13**
	e	-	-	-	-	**96.05**	**93.53**	**94.78**	**98.78**
	f	**100**	**100**	**100**	**100**	**97.35**	**96.45**	**96.90**	**98.87**
	g	**100**	**100**	**100**	**100**	91.59	**95.98**	**93.78**	**99**
	All	**100**	**100**	**100**	**100**	**73.53**	**71.33**	**72.43**	**82.93**

Chapter 6

Classification of the Thyroid Nodules

Thyroid cancer (that comes in the oral cancer category) is the second most commonly occurring cancer in the world. Ultrasound images of the thyroid nodule are often used in thyroid cancer diagnosis. As mentioned in Chapter 1, a malignant thyroid nodule produces an additional hormone called thyroxine, which causes some critical problems with patient's health and may result in his/ her death [38]. Hence, an early diagnosis of this nodule by classifying it as benign or malignant is very important.

Similar to the classification of mammogram patches, texture of a thyroid nodule, as captured in an ultrasound image, plays a crucial role in this classification [41]. Hence, we use our two earlier proposed texture-exploiting descriptors, i.e. Histogram of Oriented Texture (HOT) and Pass Band - Discrete Cosine Transform (PB-DCT) here. Since ultrasound images are large in size (200×320 after pre-processing), all extracted features may not be useful for classification. So, we again use Discrimination Potentiality (DP) to select most appropriate features. Finally, the selected features are classified using the Support Vector Machine (SVM) classifier into benign and malignant classes (binary classification). Here, we do not perform two-stage classification as done in the previous chapter because all the images present in our dataset belong to the abnormal class.

The Thyroid Digital Image Dataset (TDID) from the Universidad Nacional de

Columbia [44] is a standard database that provides ultrasound images of thyroid nodules, and most researchers test their frameworks on this dataset. hence, we apply our proposed system on all the images of TDID and compare it with the existing techniques. We achieve higher accuracy than all others (around 96%).

The rest of this chapter is organized as follows. Section 6.1 provides the summary of the related work. The proposed thyroid nodule image classification system is explained in Section 6.2. Finally, Section 6.3 presents the experimental results.

6.1 Literature Review

Previous works on the thyroid nodule classification can be grouped into two categories: deep learning-based and handcrafted-based methods [39, 38]. In the former, one constructs a learning model for feature extraction as well as classification of thyroid nodules from the captured ultrasound images. In [47], authors proposed a classification framework based upon Convolutional Neural Network (CNN). Here, they extracted the features from the input ultrasound thyroid nodule images using a pre-trained CNN, and then used SVM to classify the images into benign and malignant. The authors constructed a CNN for image classification as well where the features were extracted using a transfer learning techniques (like in VGG16-Net [46] or Inception-Net [111]). Another work that applied a transfer learning technique to classify the thyroid nodules was by Song et al. [112]. In [113], authors developed a multitask cascade convolution neural network (MC-CNN) framework to exploit the context information of thyroid nodules. Similarly, in [114], authors used deep learning via the YOLOv2 neural network to classify the thyroid nodules.

Unlike the deep learning-based methods mentioned above, the handcrafted-based methods use several traditional feature extraction techniques to extract efficient image features and a classifier to classify these features. In [41], authors used textural features like Gray Level Co-occurrence Matrix (GLCM), Gray Level Run-Length Matrix (GLRLM) and Law's texture energy measures to obtain the features. These features were then classified using the standard SVM. In [115], authors used the

Discrete Wavelet Transform (DWT) to locate the tumor region and to extract subtle information from isolated tumor region for classification. The authors in [116] performed analysis of linear and non-linear classifiers for ultrasound images. They showed that both the methods give comparable accuracy. Another study that employed handcrafted-based method was done by Raghavendra et al. [117], where they used Segmentation-based Fractal Texture Analysis (SFTA) to extract the features.

The deep learning-based methods although have high classification accuracy, they require high performance hardware as well as large amount of computational time. On the other hand, handcrafted-based methods are easy to implement without stringent hardware requirements but have low accuracy.

As mentioned earlier, in this work, we use our earlier proposed two textural descriptors (HOT and PB-DCT) with feature selection via DP for thyroid classification. Our approach belongs to the handcrafted-based methods category and overcomes the low accuracy disadvantage of these techniques. This is because a) the performance of handcrafted-based methods depends highly upon the extraction of features and little on the classifier used [38], and b) we use the texture information in a better way than other handcrafted-based methods, which also predominantly use textural properties of ultrasound images.

In Table 6.1, we summarize the strengths and weaknesses of the existing approaches as well as our proposed methods.

6.2 Classification Process

As mentioned in the previous chapter, breast cancer classification is performed on the mammogram patches that contain the cancerous tumor. These patches, also called as Region of Interests (ROIs), are identified by medical experts. However, for the thyroid nodule classification, complete unprocessed ultrasound images are available instead of these patches. Hence, a few image pre-processing steps are required before we extract the features using our proposed descriptors.

Every ultrasound thyroid image contains a background and artifacts other than the

Table 6.1: Comparison of existing and proposed classification systems.

Category	**Advantages**	**Disadvantages**
Deep Learning-based Methods	High accuracy	• High-performance hardware required • Require more processing time • There is scope for improvement
Handcrafted-based Methods	• Easy implementation • High-performance hardware not required	Low accuracy
Proposed Methods (DP-HOT and DP-PB-DCT)	• Easy implementation • High-performance hardware not required • High accuracy due to texture exploiting descriptors	DP-PB-DCT performance is poor as compared to HOT

thyroid region (see Figure 6.1a below). This background and artifact regions reduce the overall accuracy of the classification system as unwanted features get extracted due to their presence. Hence, the first step in pre-processing requires removal of these extra regions.

We use the image binarization method proposed by Otsu's et al. [118], to remove these extra regions. This method performs binarization by selecting a suitable threshold value for pixel intensity. The pixels that are darker than some threshold value are kept black, while the pixels lighter than the threshold are made white. The binarized image corresponding to input image Figure 6.1a is given in Figure 6.1b. The threshold value here is taken as 10. We can observe from Figure 6.1b that there are some bright extraneous regions detected other than the thyroid region. We simply discard these other regions and detect the region with largest size as given in Figure 6.1c. Finally, the resultant extracted thyroid region is shown in Figure 6.1d.

Once we have the extracted thyroid regions from all the images, we use the two-stage adaptive histogram equalization, as discussed earlier, to enhance them. Finally,

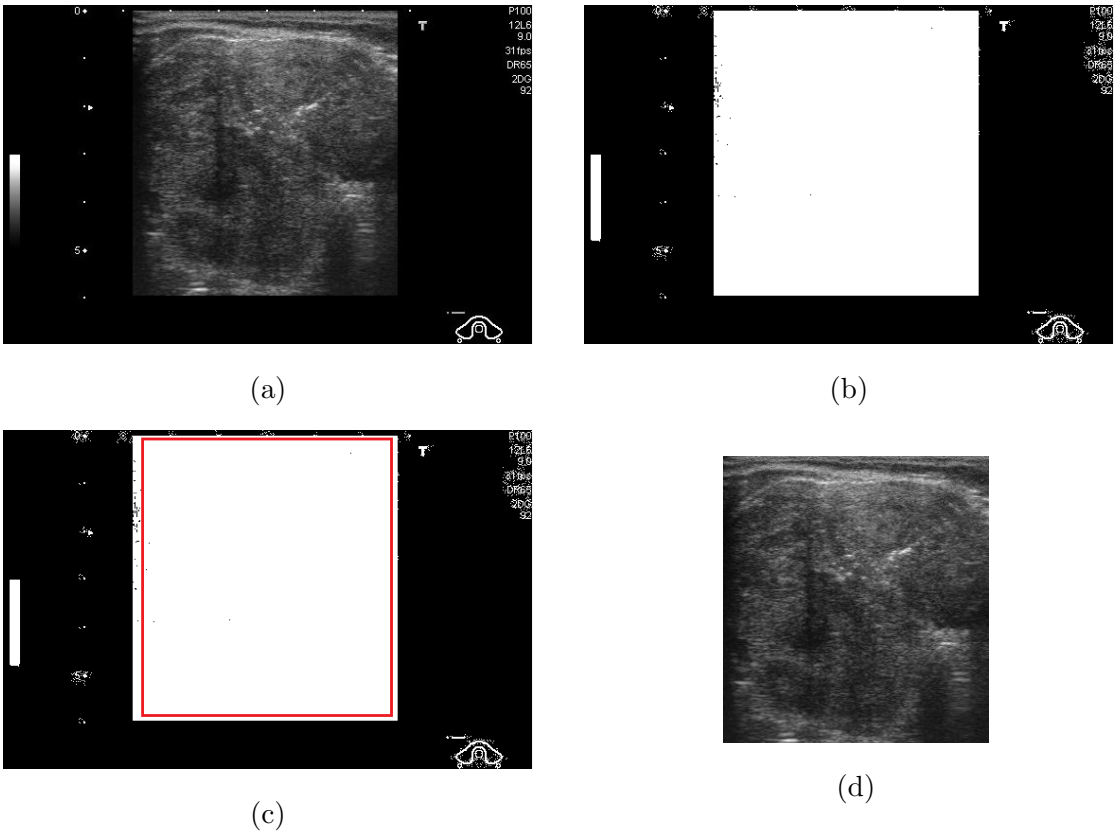


Figure 6.1: Steps in image binarization: (a) an input ultrasound thyroid image; (b) binarized image with threshold = 10; (c) largest object detection; (d) final extracted thyroid region.

we use our two earlier proposed descriptors (DP-HOT and DP-PB-DCT), to extract the features from these enhanced images.

6.3 Experimental Results

Experiments are carried out in MATLAB® 2016 on a machine with Intel i5 processor @ 2.5 GHz and 4GB RAM. As mentioned in Chapter 5, the performance of our system (and comparative systems) is evaluated by standard metrics of Sensitivity, Specificity, Accuracy, and Area Under the Curve (AUC). Again, we use a two-fold cross-validation, where the dataset is randomly divided into two equal parts. One part is used for training and the other is used for testing. We repeat two-fold cross-

validation ten times to remove any bias related to the division of the dataset.

As mentioned earlier, we use TDID [44] database for our experiments, which consists of 349 images. Each original image is of size 360×560, which becomes of size 200×320 after pre-processing. Out of these, 61 are benign, while 288 are malignant. Table 6.2 lists the number of images in TDID based on the TIRADS classes.

Table 6.2: Distribution of benign and malignant images according to the TIRADS classes.

TIRADS Class	**# of Images**	**Classification (Total Images)**
2	42	Benign
3	19	(61)
4a	96	Malignant
4b	79	(288)
4c	68	
5	45	

Since the instances of one class (benign here or minority class) are quiet less than the instances of the other class (malignant here or majority class), this dataset comes under the category of an "imbalanced" dataset [119]. Thus, in this context, many classification algorithms have low accuracy for the minority class. Most common way to solve this problem is to use Synthetic Minority Over-sampling TEchnique (SMOTE) [119]. This technique re-samples the original dataset, either by under-sampling the majority class and/ or over-sampling the minority class. Here, we perform the over-sampling of benign class so that number of instances for both the benign class and the malignant class are almost similar.

In over-sampling approach, the minority class is over-sampled by creating synthetic instances of minority class. For this, we obtain the k-nearest neighbors (from the minority class itself) for each instance of the minority class. For example, consider that one instance from the minority class has five instances in its k-nearest neighbor set.

If the amount of over-sampling required is 200% (i.e. we want to double the number of instances), then we select two neighbors from these five nearest neighbors. Subsequently, two synthetic instances are generated in the same direction as the respective neighbor and a different slope.

Finally, we compare the classification performances of our descriptors with the five existing ones mentioned earlier. The results for this are given in Table 6.3. The empty cells represent that the values are not available from their respective papers. From this table, it is evident that although DP-PB-DCT performs poorly, our DP-HOT gives almost the best results among all the descriptors.

Table 6.3: Comparison of classification accuracies for various descriptors on the TDID dataset.

Sr. No.	**Descriptors**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
1	Image Augmentation [45]	94%	**93%**	94%	-
2	VGG-16 [47]	**100%**	88%	94%	-
3	GoogLeNet [47]	-	-	79%	-
4	Circular Mask [39]	95%	64%	91%	-
5	CNN [39]	96%	66%	92%	-
6	**DP-HOT**	**100%**	90%	**96%**	**95%**
7	**DP-PB-DCT**	90%	70%	90%	88%

To summarize, we exploit the textural features of thyroid ultrasound images and use them to classify them as benign and malignant. In general, DP-HOT captures this information slightly better than DP-PB-DCT. When compared with the existing techniques, DP-HOT gives substantially better results.

Chapter 7

Conclusions and Future Work

This dissertation proposed different machine learning algorithms that focused on solving two critical real-life problems, i.e. depreciation in agricultural productivity and depleting human health. In the first-half of the dissertation, we focused on developing sampled (and hence, efficient) clustering algorithms to obtain a diverse set of plant species (or genotypes). This in turn could be used to develop better genotypes (with enhanced properties), e.g., that can be grown in less water and can survive high temperature. In the second-half of the dissertation, we presented image classification systems to accurately classify breast and thyroid cancer images as benign or malignant. Here, we developed descriptors to capture the textural properties of an image, which helped to obtain more relevant features leading to better classification.

7.1 Clustering Algorithm Variants

Variabilities in plant genotypes can be studied using their genetic and phenotypic data. Thus, Chapter 3 presented the Vector Quantized Spectral Clustering (VQSC) algorithm that is a combination of Spectral Clustering (SC) and Vector Quantization (VQ) sampling for clustering genetic data of plants. We used SC for its better clustering and VQ for its accurate sample selection. Use of this combination made our algorithm scalable for large data as well. As building the similarity matrix is critical to the SC algorithm, we exhaustively adapted four ways to build such a ma-

trix for plant genetic data. Adapting VQ for these data required using k-medoids instead of traditional k-means for finding representative samples. For a sample plant data (Soybean), we compared the performance of our VQSC algorithm with other traditional and commonly used techniques of Un-weighted Pair Graph Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ). VQSC outperformed both of these in terms of Silhouette Values (on an average 21% better than UPGMA and 24% better than NJ) and computational complexity (order of magnitude faster than both UPGMA and NJ).

Similarly, Chapter 4 presented the modified SC with Pivotal Sampling algorithm for clustering plant genotypes using their phenotypic data. Again, we used SC for its better clustering and Pivotal Sampling for its effective sample selection that in turn made our algorithm scalable for large data. For this work as well, we adapted seven different similarity measures to build the similarity matrix, which is crucial for the SC algorithm. We also presented a novel way of assigning probabilities to different genotypes for Pivotal Sampling. We performed four sets of experiments on about 2400 Soybean genotypes that demonstrated the superiority of our algorithm. First, we compare the Silhouette Values of modified SC without and with Pivotal Sampling, and show that the difference between these values is not significant. Second, when compared with the competitive clustering algorithms with samplings (i.e. SC with VQ, Hierarchical Clustering (HC) with Pivotal Sampling, and HC with VQ), Silhouette Values obtained when using our algorithm are higher. Third, our algorithm doubly outperformed the standard HC algorithm in terms of cluster quality and computational complexity (45% better clusters in terms of Silhouette Values and an order of magnitude faster than HC). Fourth and finally, we illustrate the excellence of our algorithm by comparing it with two previous works that are closest to ours.

Next, we present the future work in this context. Since the choice of the similarity matrix has a significant impact on the quality of clusters, in the future, we intend to adapt other ways of constructing this matrix such as Pearson χ^2, Squared χ^2, Bhattacharyya, Kullback-Liebler etc. [50]. Furthermore, we also plan to observe the performance of Cube Sampling, which is another probabilistic sampling technique

with data analysis properties complementary to Pivotal Sampling [5]. Both Pivotal and Cube belong to the balanced sampling category, i.e. they satisfy $Y \approx Y'_{HT}$ and $Y \approx Y'_{Hájek}$ (recall Eqs. (4.3), (4.4), and (4.5)). Cube Sampling automatically obtains the samples (without specifying the sample size), which does not happen in Pivotal. Our algorithms have been tested on the genetic and phenotypic data of Soybean plant. However, these can be applied to data of other similar plants. For example, genome sequences of Wheat, Rice, and Maize are also made from a combination of nucleotides A, T, G, and C. The only difference between Soybean sequences and sequences of these plants is the length of sequences and the numbers of Single Nucleotide Polymorphisms (SNPs) present in them, and both these things do not affect our algorithm. Similarly, phenotypic data of other plants vary only in the number of characteristics and type of characteristics, both of which again do not affect our algorithm. We have preliminarily discussed this aspect for Maize and Rice in Appendix C, with extensive experiments for these two plants planned for future [120, 64].

7.2 Cancerous Image Classification System

Early detection of cancerous tumor by classifying it as benign or malignant is important step in saving the life of the patient. Thus, Chapter 5 proposed a variant of Histogram Of Gradients (HOG) and Gabor filter combination called Histogram of Oriented Texture (HOT) for mammogram patch classification. We also revisited the Pass Band - Discrete Cosine Transform (PB-DCT) descriptor for the same. We used the feature selection technique of Discrimination Potentiality (DP) with the above two descriptors for reduction in feature space. This resulted in two new descriptors (DP-HOT and DP-PB-DCT). We considered the density of mammogram patches as a factor for classification (this was not done earlier), and showed that this plays an important role in classification.

We tested our two-stage mammogram patch classification system (normal–abnormal and benign–malignant), using the two new descriptors for each density class, on all the images of the MIAS and DDSM datasets from the Image Retrieval in Medical

Application (IRMA) repository (in literature, experiments had been done only on a subset of these images). This helped achieve a high classification performance (in terms of specificity, sensitivity, accuracy and AUC) in an absolute sense as well as in relative sense (compared with the six standard descriptors). We achieved an average accuracy of more than 92%, which turned out to be categorically more than all of the existing standard descriptors. Our descriptors captured the textural information in a mammogram patch well, which led to this improvement.

Similar to the mammogram patch classification, the texture of a thyroid nodule also plays a vital role in classification of the nodule as benign or malignant. Therefore, Chapter 6 proposed a thyroid image classification method by applying the above-mentioned texture exploiting descriptors. Apart from pre-processing the images for illumination normalization and visibility enhancement, we also applied image binarization to remove the background and artifact regions from the thyroid ultrasound images. We used the Thyroid Digital Image Dataset (TDID) for our experiments. Results showed that although our system performed poorly with DP-PB-DCT, we achieved a high average classification accuracy with DP-HOT (around 96%; substantially more than multiple previously proposed methods).

The *first* future direction here involves developing a Computer-Aided Diagnosis (CAD) application for breast and thyroid cancer diagnoses using our framework. This will aid doctors (radiologists) reach more accurate results. Apart from the diagnosis, the prognoses of the cancer is equally difficult. This includes predicting the further development of the cancer [121]. Although a large number of research articles are available for diagnoses of different cancers, very few talk about their prognoses [122, 123]. Thus, the *second* future direction is to develop prognostic classification models using machine learning algorithms. These models could save the patients from receiving superfluous treatment and its associated medical cost. Furthermore, lung cancer is another commonly prevailing cancer whose cases are increasing at an alarming rate. Hence, *third*, we plan to develop an accurate feature extraction technique that can obtain the features from the computed tomography images of the lung cancer patient [124]. *Fourth*, we plan to perform two-stage classification for thyroid ultrasound as

well. Here, we wish to classify the benign images into the TIRADS score of 2 & 3, and malignant images into TIRADS score of 4 & 5.

Finally, we present the future work that aims to bring the two application areas together. Since both clustering and classification are machine learning techniques to group similar data, there is always a scope that we can use these techniques for any application area. For example, in the cancer image classification context, we plan to develop a clustering technique that can be used to obtain the crucial features from all the available features. Here, we intend to cluster all the features and select that cluster, which contains the most important features that can be used to distinguish the two classes. Similarly, classification techniques can be applied in the plant domain, where we obtain the features from several images of the given plant and decide whether they have a disease or not. Here, we plan to build the classification model that would be trained on several healthy and diseased plant images. Then, we can use this model to predict the unknown images of the plant.

Bibliography

[1] E. Vogel, M. Donat, L. Alexander, et al. The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, 14(5):054010, 2019.

[2] The Hindu. One in 10 Indians will develop cancer during their lifetime: WHO report. `https://www.thehindu.com/sci-tech/health/one-in-10-indians-will-develop-cancer-during-their-lifetime-who-report/article30734480.ece`. Accessed: October 2020.

[3] A. Fahad, N. Alshatri, Z. Tari, et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, 2014.

[4] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14:849–856, 2001.

[5] Y. Tille. *Sampling Algorithms*. Springer, Springer-Verlag New York, 2006.

[6] A. Friedrich, R. Ripp, N. Garnier, et al. Blast sampling for structural and functional analyses. *BMC Bioinformatics*, 8(1):1–17, 2007.

[7] X. Wang, X. Zheng, F. Qin, et al. A fast spectral clustering method based on growing vector quantization for large data sets. In *Proceedings of International Conference on Advanced Data Mining and Applications*, pages 25–33, 2013.

[8] T. Backeljau, L. Bruyn, H. Wolf, et al. Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. *Molecular Biology and Evolution*, 13:309–313, 1996.

[9] T. Lee, H. Guo, X. Wang, et al. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, 15:162, 2014.

[10] P. Painkra, R. Shrivatava, S. Nag, et al. Clustering analysis of Soybean germplasm (Glycine max L. Merrill). *The Pharma Innovation Journal*, 7(4):781–786, 2018.

[11] P. Ingvarsson and N. Street. Association genetics of complex traits in plants. *New Phytologist*, 189(4):909–922, 2011.

[12] Z. Zhou, Y. Jiang, Z. Wang, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in Soybean. *Nature Biotechnology*, 33(4):408–414, 2015.

[13] S. Subramanian, U. Ramasamy, and D. Chen. VCF2PopTree: a client-side software to construct population phylogeny from genome-wide SNPs. *PeerJ*, 7:e8213, 2019.

[14] P. Sharma, S. Sareen, M. Saini, et al. Assessing genetic variation for heat tolerance in synthetic wheat lines using phenotypic data and molecular markers. *Australian Journal of Crop Science*, 8(4):515–522, 2014.

[15] A. Kahraman, M. Onder, and E. Ceyhan. Cluster analysis in common bean genotypes (Phaseolus vulgaris L.). *Türkish Journal of Agricultural and Natural Sciences*, 1:1030–1035, 2014.

[16] D. Mullner. fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–8, 2013.

[17] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[18] G. Chauvet. On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4):1320–1340, 2012.

[19] C. Gireesh, S. Husain, M. Shivakumar, et al. Integrating principal component score strategy with power core method for development of core collection in Indian Soybean germplasm. *Plant Genetic Resources*, 15(3):230–238, 2015.

[20] Sumeet Shah. Statistics of breast cancer in India. `http://www.breastcancerindia.net/statistics/stat_global.html`. Accessed: March 2016.

[21] A. Oliver, J. Freixenet, J. Marti, et al. A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, 14(2):87–110, 2010.

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

[23] S. Anand and S. Gayathri. Mammogram image enhancement by two-stage adaptive histogram equalization. *Optik-International Journal for Light and Electron Optics*, 126(21):3150–3152, 2015.

[24] S. Jenifer, S. Parasuraman, and A. Kadirvelu. Contrast enhancement and brightness preserving of digital mammograms using fuzzy clipped contrast-limited adaptive. *Applied Soft Computing*, 42:167–177, 2016.

[25] M. Sundaram, K. Ramar, N. Arumugam, et al. Histogram modified local contrast enhancement for mammogram images. *Applied Soft Computing*, 11(8):5809–5816, 2011.

[26] N. Mudigonda, R. Rangayyan, and J. Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging*, 19(10):1032–1043, 2000.

[27] A. Tahmasbi, F. Saki, and S. Shokouhi. Classification of benign and malignant masses based on zernike moments. *Computers in Biology and Medicine*, 41(8):726–735, 2011.

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

[29] L. Nanni, S. Brahnam, S. Ghidoni, et al. Different approaches for extracting information from the co-occurrence matrix. *PLOS ONE*, 8(12):e83554, 2013.

[30] S. Beura, B. Majhi, and R. Dash. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. *Neurocomputing*, 154:1–14, 2015.

[31] S. Ergin and O. Kilinc. A new feature extraction framework based on wavelets for breast cancer diagnosis. *Computers in Biology and Medicine*, 51:171–182, 2015.

[32] J. De Oliveira, A. De Albuquerque, and T. Deserno. Content-based image retrieval applied to BIRADS tissue classification in screening mammography. *World Journal of Radiology*, 3(1):24–31, 2011.

[33] A. Oliver, J. Freixenet, R. Martí, et al. A novel breast tissue density classification methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65, 2008.

[34] A. Oliver, A. Torrent, X. Lladó, et al. Automatic microcalcification and cluster detection for digital and digitised mammograms. *Knowledge Based Systems*, 28:68–75, 2012.

[35] S. Petroudi and M. Brady. Breast density characterization using texton distributions. In *Proceedings of International Conference on Engineering in Medicine and Biology Society*, pages 5004–5007, 2011.

[36] T. Deserno. IRMA database. http://ganymed.imib.rwth-aachen.de/deserno/datasets_en.php?SELECTED=00014#00014.dataset, 2012.

[37] American Society of Clinical Oncology. Thyroid cancer: Diagnosis. https://www.cancer.net/cancer-types/thyroid-cancer/diagnosis. Accessed: October 2020.

[38] D. Nguyen, T. Pham, G. Batchuluun, et al. Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains. *Journal of Clinical Medicine*, 8(11):1976, 2019.

[39] D. Nguyen, J. Kang, T. Pham, et al. Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *Sensors*, 20(7):1822, 2020.

[40] J. Chi, E. Walia, P. Babyn, et al. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of Digital Imaging*, 30(4):477–486, 2017.

[41] C. Chang, S. Chen, and M. Tsai. Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. *Pattern Recognition*, 43(10):3494–3506, 2010.

[42] D. Gaitini, R. Evans, and G. Ivanac. Chapter 16: thyroid ultrasound. *EFSUMB Course Book*, 2011.

[43] J. Kwak, K. Han, J. Yoon, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology*, 260(3):892–899, 2011.

[44] L. Pedraza, C. Vargas, F. Narváez, et al. An open access thyroid ultrasound image database. In *Proceedings of International Symposium on Medical Information Processing and Analysis*, page 92870W, 2015.

[45] Y. Zhu, Z. Fu, and J. Fei. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning. In *Proceedings*

of IEEE International Conference on Computer and Communications, pages 1819–1823, 2017.

[46] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] K. Sundar, K. Rajamani, and S. Sai. Exploring image classification of thyroid ultrasound images using deep learning. In *Proceedings of International Conference on ISMAC in Computational Vision and Bio-Engineering*, pages 1635–1641, 2018.

[48] S. Kotsiantis, I. Zaharakis, and P. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[49] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, 22(8):888–905, 2000.

[50] S. Cha. Comprehensive survey on distance/ similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 4(1):300–307, 2007.

[51] A. Vasuki and P. Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47, 2006.

[52] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[53] J. Deville and Y. Tille. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101, 1998.

[54] S. Haykin. *Neural Networks and Learning Machines*. Pearson Education India, 2010.

[55] A. Vignal, D. Milan, M. SanCristobal, et al. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34(3):275–305, 2002.

[56] M. Bouaziz, C. Paccard, M. Guedj, et al. SHIPS: spectral hierarchical clustering for the inference of population structure in genetic studies. *PLoS ONE*, 7(10):e45685, 2012.

[57] J. Felsenstein. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46:101–111, 1997.

[58] T. Jukes and C. Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132, 1969.

[59] D. Betel, M. Wilson, A. Gabow, et al. The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(1):149–153, 2008.

[60] D. Lawson and D. Falush. Similarity matrices and clustering algorithms for population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13:337–361, 2012.

[61] L. Li, M. Shiga, W. Ching, et al. Annotating gene functions with integrative spectral clustering on microarray expressions and sequences. *Genome Informatics*, 22:95–120, 2010.

[62] J. Zhang, A. Mamlouk, T. Martinetz, et al. PhyloMap: an algorithm for visualizing relationships of large sequence data sets and its application to the influenza A virus genome. *BMC Bioinformatics*, 12(1):1–19, 2011.

[63] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[64] S. Immanuel, N. Pothiraj, K. Thiyagarajan, et al. Genetic parameters of variability, correlation and path-coefficient studies for grain yield and other yield

attributes among rice blast disease resistant genotypes of rice (Oryza sativa L.). *African Journal of Biotechnology*, 10(17):3322–3334, 2011.

[65] B. Divya, S. Robin, A. Biswas, et al. Genetics of association among yield and blast resistance traits in rice (Oryza sativa). *Indian Journal of Agricultural Sciences*, 85(3):354–360, 2015.

[66] F. Huang, Y. Gan, D. Zhang, et al. Leaf shape variation and its correlation to phenotypic traits of Soybean in northeast China. In *Proceedings of International Conference on Bioinformatics and Computational Biology*, pages 40–45, 2018.

[67] V. Carpentieri-Pipolo, K. de Almeida Lopes, and G. Degrassi. Phenotypic and genotypic characterization of endophytic bacteria associated with transgenic and non-transgenic Soybean plants. *Archives of Microbiology*, 201(8):1029–1045, 2019.

[68] S. Islam, J. Anothai, C. Nualsri, et al. Genetic variability and cluster analysis for phenological traits of Thai Indigenous Upland Rice (Oryza sativa L.). *Indian Journal of Agricultural Research*, 54(2):211–216, 2020.

[69] H. Fried, S. Narayanan, and B. Fallen. Characterization of a Soybean (Glycine max L. Merr.) germplasm collection for root traits. *PLoS ONE*, 13(7):e0200463, 2018.

[70] A. Stansluos, A. Öztürk, S. Kodaz, et al. Genetic diversity in sweet corn (Zea mays L. saccharata) cultivars evaluated by agronomic traits. *Mysore Journal of Agricultural Sciences*, 53(1):1–8, 2019.

[71] J. Hancock and T. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7:1–41, 2020.

[72] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[73] S. Srisuk and A. Petpon. A gabor quotient image for face recognition under varying illumination. In *Proceedings of International Symposium on Visual Computing*, pages 511–520, 2008.

[74] W. Kong, C. Sun, S. Hu, et al. Automatic spectral clustering and its application. In *Proceedings of International Conference on Intelligent Computation Technology and Automation*, pages 841–845, 2010.

[75] A. Rutherford. *ANOVA and ANCOVA: a GLM approach.* Wiley, 2011.

[76] W. Beyer. *Handbook of tables for probability and statistics.* CRC Press, 2019.

[77] D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[78] J. Hájek. Comment on "an essay on the logical foundations of survey sampling, part one". In V. Godambe and D. Sprott, editors, *The Foundations of Survey Sampling*. Holt, Rinehart and Winston, Toronto, 1971.

[79] J. Kamarainen, V. Kyrki, and H. Kalviainen. Invariance properties of gabor filter-based features - overview and applications. *IEEE Transactions on Image Processing*, 15(5):1088–1099, 2006.

[80] S. Beura, B. Majhi, and R. Dash. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. *Neurocomputing*, 154:1–14, 2015.

[81] W. Peng, R. Mayorga, and E. Hussein. An automated confirmatory system for analysis of mammograms. *Computer Methods and Programs in Biomedicine*, 125:134–144, 2016.

[82] G. Rabottino, A. Mencattini, M. Salmeri, et al. Mass contour extraction in mammographic images for breast cancer identification. In *Proceedings of IMEKO*

TC4 Symposium - Exploring New Frontiers of Instrumentation and Methods for Electrical and Electronic Measurements, 2008.

[83] S. Shanthi and V. Bhaskaran. Computer aided detection and classification of mammogram using self-adaptive resource allocation network classifier. In *Proceedings of International Conference on Pattern Recognition, Informatics and Medical Engineering*, pages 284–289, 2012.

[84] M. Abdel-Nasser, H. Rashwan, D. Puig, et al. Analysis of tissue abnormality and breast density in mammographic images using a uniform local directional pattern. *Expert Systems with Applications*, 42(24):9499–9511, 2015.

[85] I. Buciu and A. Gacsadi. Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control*, 6(4):370–378, 2011.

[86] N. Gedik. A new feature extraction method based on multi-resolution representations of mammograms. *Applied Soft Computing*, 44:128–133, 2016.

[87] J. Leena Jasmine, A. Govardhan, and S. Baskaran. Microcalcification detection in digital mammograms based on wavelet analysis and neural networks. In *Proceedings of International Conference on Control, Automation, Communication and Energy Conservation*, pages 1–6, 2009.

[88] S. Dabbaghchian, M. Ghaemmaghami, and A. Aghagolzadeh. Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology. *Pattern Recognition*, 43(4):1431–1440, 2010.

[89] M. Laadjel, S. Al-Maadeed, and A. Bouridane. Combining Fisher locality preserving projections and passband DCT for efficient palmprint recognition. *Neurocomputing*, 152:179–189, 2015.

[90] C. Muramatsu, T. Hara, T. Endo, et al. Breast mass classification on mammograms using radial local ternary patterns. *Computers in Biology and Medicine*, 72:43–53, 2016.

[91] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Proceedings of International Conference on Image and Signal Processing*, pages 236–243, 2008.

[92] A. Oliver, X. Lladó, J. Freixenet, et al. False positive reduction in mammographic mass detection using local binary patterns. In *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 286–293, 2007.

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

[94] S. Wajid and A. Hussain. Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. *Expert Systems with Applications*, 42(20):6990–6999, 2015.

[95] N. Mudigonda, R. Rangayyan, and J. Desautels. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Transactions on Medical Imaging*, 20(12):1215–1227, 2001.

[96] S. Khan, M. Hussain, H. Aboalsamh, et al. Optimized gabor features for mass classification in mammography. *Applied Soft Computing*, 44:267–280, 2016.

[97] C. Conde, D. Moctezuma, I. De Diego, et al. HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments. *Neurocomputing*, 100:19–30, 2013.

[98] H. Ouanan, M. Ouanan, and B. Aksasse. Gabor-HOG features based face recognition scheme. *Indonesian Journal of Electrical Engineering and Computer Science*, 15(2):331–335, 2015.

[99] X. Xu, C. Quan, and F. Ren. Facial expression recognition based on Gabor Wavelet transform and Histogram of Oriented Gradients. In *Proceedings of*

International Conference on Mechatronics and Automation, pages 2117–2122, 2015.

[100] J. De Oliveira, T. Deserno, and A. De Albuquerque. Breast lesions classification applied to a reference database. In *Proceedings of International Conference on E-Medical Systems*, pages 29–31, 2008.

[101] T. Deserno, M. Soiron, and J. De Oliveira. Texture patterns extracted from digitizes mammograms of different BI-RADS classes. *Image Retrieval in Medical Applications Project*, release 1, 2012.

[102] K. Panetta, Y. Zhou, S. Agaian, et al. Nonlinear unsharp masking for mammogram enhancement. *IEEE Transactions on Information Technology in Biomedicine*, 15(6):918–928, 2011.

[103] J. Kamarainen, V. Kyrki, and H. Kälviäinen. Invariance properties of gabor filter-based features - overview and applications. *IEEE Transactions on Image Processing*, 15(5):1088–1099, 2006.

[104] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60, 2002.

[105] W. Kao, M. Hsu, and Y. Yang. Local contrast enhancement and adaptive feature extraction for illumination-invariant face recognition. *Pattern Recognition*, 43(5):1736–1747, 2010.

[106] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[107] Y. Zeng, J. Luo, and S. Lin. Classification using markov blanket for feature selection. In *Proceedings of International Conference on Granular Computing*, pages 743–747, 2009.

[108] A. Dong and B. Wang. Feature selection and analysis on mammogram classification. In *Proceedings of Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 731–735, 2009.

[109] R. Nandi, A. Nandi, R. Rangayyan, et al. Genetic programming and feature selection for classification of breast masses in mammograms. In *Proceedings of International Conference on Engineering in Medicine and Biology Society*, pages 3021–3024, 2006.

[110] S. Yeh. Using trapezoidal rule for the area under a curve calculation. In *Proceedings of Annual SAS® User Group International*, 2002.

[111] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[112] J. Song, Y. Chai, H. Masuoka, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine*, 98(15):e15133, 2019.

[113] W. Song, S. Li, J. Liu, et al. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1215–1224, 2018.

[114] L. Wang, S. Yang, S. Yang, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World Journal of Surgical Oncology*, 17(1):1–9, 2019.

[115] V. Sudarshan, M. Mookiah, U. Acharya, et al. Application of wavelet techniques for cancer diagnosis using ultrasound images: A review. *Computers in Biology and Medicine*, 69:97–111, 2016.

[116] F. Ouyang, B. Guo, L. Ouyang, et al. Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. *European Journal of Radiology*, 113:251–257, 2019.

[117] U. Raghavendra, U. Acharya, A. Gudigar, et al. Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions. *Ultrasonics*, 77:110–120, 2017.

[118] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[119] N. Chawla, K. Bowyer, L. Hall, et al. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[120] G. Galli, D. Lyra, F. Alves, et al. Impact of phenotypic correction method and missing phenotypic data on genomic prediction of maize hybrids. *Crop Science*, 58(4):1481–1491, 2018.

[121] S. Sayed. Machine learning is the future of cancer prediction. `https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa`. Accessed: January 2021.

[122] P. Ferroni, F. Zanzotto, S. Riondino, et al. Breast cancer prognosis using a machine learning approach. *Cancers*, 11(3):328, 2019.

[123] D. Sun, M. Wang, and A. Li. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3):841–850, 2018.

[124] J. Kuruvilla and K. Gunavathi. Lung cancer classification using neural networks for CT images. *Computer Methods and Programs in Biomedicine*, 113(1):202–209, 2014.

[125] N. Belalia, A. Lupini, A. Djemel, et al. Analysis of genetic diversity and population structure in Saharan maize (Zea mays L.) populations using phenotypic traits and SSR markers. *Genetic Resources and Crop Evolution*, 66(1):243–257, 2019.

Appendix A

Validation of Soybean Phenotypic Data

Here, we first present phenotypic data of the Soybean genotypes used for our experiments in Chapter 4. Please see Table A.1 below. As mentioned earlier, EPV: Early Plant Vigor, PH: Plant Height, NPB: Number of Primary Branches, LS: Lodging Score, NPPP: Number of Pods Per Plant, SW: 100 Seed Weight, SYPP: Seed Yield Per Plant, and Days to Pod Initiation (DPI).

Table A.1: Phenotypic data of the Soybean genotypes used for experiments.

Genotypes	**EPV**	**PH**	**NPB**	**LS**	**NPPP**	**SW**	**SYPP**	**DPI**
1	Poor	54	6.8	Moderate	59.8	6.5	2.5	65
2	Poor	67	3.4	Severe	33	6.2	3.9	64
3	Good	38.4	2.8	Slight	68	6.9	4.4	61
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Very Good	89.6	5	Severe	32.6	7.3	3.4	62

Next, we validate this data. For this, we compare our phenotypic data with a similar Soybean phenotypic data from [19] for the common set of phenotypic characteristics (PH, NPPP, DPI). For comparison purpose, we work with original (non-normalize)

values of these characteristics. This comparison is done using standard statistical metrics and is given in Table A.2 below.

Table A.2: Comparison of SD, CV, mean, and range for our phenotypic data and similar previous data.

Parameter	**Work**	**PH**	**NPPP**	**DPI**
Standard Deviation (SD)	Our Work	16.61	20.16	7.85
	Previous Work [19]	18.6	24.1	8
Coefficient of Variance (CV)	Our Work	31.80	47.13	13.62
	Previous Work [19]	30.9	55.2	17.8
Mean	Our Work	52.24	42.78	57.60
	Previous Work [19]	60.3	43.6	54.7
Range	Our Work	13-102	4.33-197.66	24-80
	Previous Work [19]	5.4-118.8	1.33-301	30-98

From this table, it is evident that the Standard Deviation (SD), Coefficient of Variance (CV), and Mean of our data and the data from the previous work are very close (for all three characteristics of PH, NPPP, and DPI). The slight variation in the metrics between the two data for all the characteristics is due to the difference in the ranges of the respective characteristics (due to the slightly differing selection of the genotypes by the two works).

Appendix B

Comparison of Pivotal Sampling with Other Samplings

Here, we compare the Pivotal Sampling technique discussed in Chapter 4 with those proposed by Gireesh et al. [19] for a similar dataset. As earlier, we do sampling on 2376 Soybean genotypes while Gireesh et al. performed the Principal Component Score (PCS) and the Power Core (PC) samplings on 3443 Soybean genotypes. Since the samples obtained by the PC method are better, we compare our results with this sampling only.

This comparison is done using the statistical metrics of Standard Deviation (SD), Coefficient of Variance (CV) and Mean, and is given in Table B.1 below. In the table, PH: Plant Height, NPPP: Number of Pods Per Plant, DPI: Days to Pod Initiation. Again, for comparison, we work with original (non-normalize) values of the characteristics. Since the metrics of our sampled data are more closer to our respective full data as compared to the metrics of the previous works' sampled data to its respective full data, our sampling is better.

Table B.1: Comparison of Pivotal Sampling and Power Core method for three characteristics.

Parameters	**Work**	**Population**	**PH**	**NPPP**	**DPI**
Standard Deviation (SD)	Our Work	Overall	16.61	20.16	7.85
		Sampled	17.34	18.90	7.42
	Previous Work [19]	Overall	18.6	24.1	8
		Sampled	22.15	45.33	11.73
Coefficient of Variance (CV)	Our Work	Overall	31.80	47.13	13.62
		Sampled	31.91	43.97	13.03
	Previous Work [19]	Overall	30.9	55.2	17.8
		Sampled	39.86	91.06	25.46
Mean	Our Work	Overall	52.24	42.78	57.60
		Sampled	54.34	42.99	56.94
	Previous Work [19]	Overall	60.3	43.6	54.7
		Sampled	55.57	49.78	56.65

Appendix C

Modified Spectral Clustering for Maize and Rice Phenotypic Data

In Chapter 4, we have demonstrated the usefulness of our proposed algorithm (modified Spectral Clustering (SC) with Pivotal Sampling) on the genotypes of the Soybean plant. Here, we demonstrate our algorithms' applicability to the genotypes of the other two plants (Maize and Rice). The phenotypic data for the Maize genotypes is given in Table C.1 [125], and for the Rice genotypes is given in Table C.2 [64, 68]. In the tables, DS: Days to Silking, PH: Plant Height, EH: Ear Height, ED: Ear Diameter, EL: Ear Length, SW: 100 Seed Weight, TN: Tiller Number, PN: Panicle Number, PL: Panicle Length, BDR: Blast Disease Resistance.

Table C.1: Phenotypic data of the Maize genotypes.

Genotypes	**DS**	**PH**	**EH**	**ED**	**EL**	**SW**
1	77	75	33	3.2	11.6	2.3
2	98	45	14	2.7	8.1	1.6
3	68	132	80	3.7	16.2	3.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	70	50	35	3.1	10.6	2.6

Table C.2: Phenotypic data of the Rice genotypes.

Genotypes	**TN**	**PH**	**PN**	**PL**	**SW**	**BDR**
1	6.8	124.2	5.5	25.6	22.1	Resistant
2	6.5	121.6	6.8	24.8	23.1	Moderately Resistant
3	7.2	126.4	4.5	26.1	19.5	Moderately Susceptible
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	7.1	131.4	5.1	25.9	18.5	Susceptible

We can observe from Tables A.1, C.1, and C.2 that there is a set of common phenotypic characteristics for the three plant genotypes. Also, the values of all the characteristics are either categorical or numerical. As mentioned earlier, the categorical values can be easily converted to numerical ones. Since the input to our algorithm is a matrix built using the phenotypic data for given genotypes, it can be applied to any of these plants.

To demonstrate the usefulness of our algorithm to the two new plant genotypes, without loss of generality, we perform clustering of Rice genotypes using our modified SC[1]. For this, we use the data from Islam et al. [68], where the authors have used Hierarchical Clustering (HC) to cluster ten Rice genotypes into three clusters. Hence, we also cluster these ten genotypes into three clusters using our modified SC. In [68], the output is in the form of a hierarchical tree, which is non-numerical, and hence, difficult to compare. Thus, we compute Silhouette Values for our modified SC and HC. This data for the four similarity measures are given in Table C.3. As evident from this table, our algorithm substantially outperforms HC.

[1]Recall that Islam et al. does not perform any sampling. Hence, for fair comparison, we also perform only clustering and not sampling.

Table C.3: Silhouette Values of modified SC and HC for three clusters of ten Rice genotypes.

Similarity Measure	**modified SC**	**HC**
Euclidean	0.2743	0.0076
Squared Euclidean	0.3276	0.0253
City-block	0.2561	0.0219
Correlation	0.3265	0.0433