Prostate Cancer Analysis Using MapReduce and Unsupervised Learning Methods

Zoya Yasin

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Professor Khurshid Ahmad

August 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Zoya Yasin

August 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Zoya Yasin

August 19, 2022

Prostate Cancer Analysis Using MapReduce and Unsupervised Learning Methods

Zoya Yasin, Master of Science in Computer Science
University of Dublin, Trinity College, 2022

Supervisor: Professor Khurshid Ahmad

In recent times there have been many advancements in the field of digital pathology, owing to the use of Whole Slide Imaging (WSI) technology and the availability of high-resolution imaging sensors. However, there is still a lot that needs to be done to revolutionize diagnosis. There has been a growing demand for faster and more affordable diagnoses of chronic diseases like cancer, and the field of medicine could benefit from the use of automated systems that aid the screening and evaluation of tissue samples taken from patients. Additionally, the general trends in pathological analysis favour shifting to a digital platform, but the overwhelming size of datasets created by the generation of high-resolution medical images is a huge deterrent. In accordance with said requirements, this study has proposed an approach to medical image analysis with the help of a parallel computing framework (MapReduce) and unsupervised clustering methods. The experiments in the proposed system utilize a dataset of prostate tissue images. Due to parallel processing abilities, this system can process and extract quantifiable features from the image in significantly smaller amounts of time than traditional systems. The features extracted using the above framework are robust to changes in positioning and noise and eliminate redundancy of information presented by them. The system is successfully able to distinguish between two separate categories of prostate tissue based on learning from the extracted feature vector.

Acknowledgments

I would like to offer my sincerest thanks to my Supervisor, Professor Khurshid Ahmad, without whose expert guidance and insightful advice this dissertation would not have been a successful endeavor. His vast knowledge in multiple disciplines and patience in explaining concepts made this research undertaking a very rewarding one.

The implementation of the methods proposed in this work would not have been possible without the support of Dr. Aamir Ahmad, who provided me with the dataset used throughout the research. I am extremely grateful for his efforts involved.

I would also like to thank Akash Verma, alumni of Trinity College Dublin, who provided me with useful tips and insights based on his previous research and helped me gain a perspective on the problem domain.

I am very grateful for the support of my friends Shriya Vikhram and Shubham Uniyal, they have been a constant source of motivation during the course of this research.

Lastly, I would like to thank my family for their encouragement to pursue my dreams and providing moral and emotional support in times of need. They are the ones who made all my endeavors in life possible and I cannot express enough gratitude for everything they have provided me with.

Zoya Yasin

University of Dublin, Trinity College
August 2022

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We live in a world which generates an overwhelmingly large amount of data every day, data which must be stored, processed, analysed and have information extracted from them. Digital images happen to be one of the primary types of data generated, their popularity compounded by the abundance of imaging devices available today. Additionally, the fact that they are an ideal format for storing information that the human brain can process efficiently make images an accepted format of knowledge representation. Visual perception is one of the senses that humans rely on most - one glance at an image, and the brain registers multiple descriptors and facets of the information contained within. It is no wonder then that a considerable effort was put into improving the digitisation process and respective analysis of images - the applications of which are seen everywhere today, be it in the field of arts, engineering or medicine. Today, image processing techniques are being used for capturing data that cannot be processed by human eyes, gathering scientific knowledge, restoration purposes and medical diagnostics. Moving images to the digital platform has enabled humankind to take leaps that were earlier inconceivable.

As with any other process being carried out on an industrial scale, automation has become a necessity in the field of image analysis as well. The huge volumes of imaging data being generated keeps adding on to the issues related to storage and retrieval of these files, and the manual analysis of these images, if required, proves to be tedious. The human visual system is prone to reporting erroneous knowledge when exhausted. This has led to the development of systems that aid in the visual analysis of images - softwares like Google Lens, developed by Google, aids users in the identification of everyday objects they encounter and come up with relevant search results. There are several benefits to automating the image analysis process - a big one being that systems built for these purposes consistently perform to the best of their ability, unlike their human counterparts. Additional benefits include better and more efficient storage/ retrieval systems and cutting down the costs involved in more traditional methods of analysis.

As mentioned previously, image analysis has also found its use in the medical domain, particularly for pathological analysis of samples taken from patients. Expert pathologists are able to assess certain attributes like structure, patterns, hue and texture of the sample taken from an individual and then diagnose them. The digital format of glass slide images enables them to analyse the sample in a much more convenient and effective manner by helping them focus on the region of interest and achieve better magnification. The analysis of pathological images can be divided into several stages - (1) obtaining high-resolution images of samples, (2) extracting relevant features from each of the images and then (3) using these features to diagnose and evaluate the health conditions of the patient. Obtaining these images is not an extremely challenging task in today's day and age - the advent of Whole Slide Imaging (WSI) technology has enabled professionals to take scans of entire slides of biological samples and convert them into high-resolution images. However, a significant roadblock in the efforts being made to digitise pathological analysis seems to be the processing required to extract features from the images. Both the choice of descriptors and tools used for this process are of critical importance, and currently available technology is either computationally expensive, time-consuming, inaccurate or both.

There has been substantial research in the attempt to automate the diagnosis of medical conditions through the help of image data and machine learning algorithms as well. Systems have been developed using the concepts in these research works, and they have shown significantly good results. However, there is not much room for error in the field of pathology. A single misdiagnosed case is equivalent to a lost life, making it an error that can't be undone. As such, automated systems in medical image analysis can be used for screening purposes at best, leaving the diagnosis to human experts. Time is also of the essence in this whole process, and without efficient systems for the storage and processing of these images for extraction of features, it is unrealistic to expect that computer-based systems will ever replace traditional human analysis in the field of medicine.

Therefore, this work has attempted to provide solutions to the problems mentioned above based on the available work done for managing Big Data and machine learning-based classification/clustering approaches to analyse the data at hand. This work gives an insight into the current state-of-the-art of the methods used in the proposed system and builds upon them to develop a more efficient and robust model.

1.1 Problem Definition

As outlined in the previous section, several problems emerge when discussing the digitisation of medical image analysis. The pipeline of medical image analysis can be divided into the following four stages:

1. **Image Acquisition** - Involves the collection of medical tissue or cell samples, followed by the process of digitisation using WSI technology.

2. **Image Storage and Retrieval** - Using storage systems and frameworks for the storage of large datasets of high-resolution images and efficient retrieval.

3. **Feature Extraction** - The process of finding quantifiable (or other) attributes and extracting them from the image for a better description of each sample.

4. **Analysis and Diagnosis** - The final stage, where the extracted features are examined to find any abnormalities in structure or appearance, and a respective diagnosis is suggested.

All these stages require a lot of effort and research to better understand and implement a computer-based system that facilitates automated screening and diagnosis. The acquisition process will always require some human intervention - after all, bodily samples from a living individual are being collected. The storage and retrieval process can be handled by modern Big Data management systems - there are several alternatives available out there today. However, things start getting more complicated when one arrives at the feature extraction and analysis stages. Several questions arise when choosing the approach to both these steps:

1. What is the best way to represent an image so that it encompasses all the relevant features of the sample in a machine-readable format?

2. What methods can be utilised to extract said features from the images in the most economical and time-efficient manner?

3. How can the extracted features be analysed in an efficient manner to distinguish between healthy and sick samples?

For the representation of features in tissue and cell images, it makes sense to segment images and extract the region of interest. This could include the tissue formation patterns, nuclei of the cell, lobule structure, etc. However, a system focusing on these types of features will have different input requirements for different types of samples, and the solution will not be generic. The structural makeup of the sample can also be examined, but it is important to consider the colour and hue information encoded in the image for better understanding.

Even if the optimal choice of descriptors is identified, extracting them as a numerical vector for classification models to process can be a time-intensive task. The use of such descriptors requires lengthy calculations to be performed, which are also computationally expensive. Given that medical images are large files with sizes in the range of terabytes, the data being processed is so huge that they add on to the delay in processing times. This is undesirable in the medical domain as diagnosis needs to be fast and affordable for the general benefit of the public.

As for the third aspect of this problem, the choice of the type of classifier is critical in deciding the accuracy of the system. Given the right set of features, a classification system can help in making the right distinction between healthy and sick samples. However, the model chosen must be a right fit for the data at hand and the requirements of the system. The samples collected initially are never labelled as they have not undergone evaluation by an expert. While supervised methods of clustering are preferred for the ease of training and performance assessment, they wouldn't exactly be suitable for working with unlabelled data. Unsupervised methods are best suited for these purposes, but their evaluation is slightly more complex.

1.2 Contributions

The key contribution of this work is an approach to medical image analysis using a parallel computing framework and clustering methods. The proposed approach allows for the streamlined extraction of a global set of features from the images and then utilises these features to analyse the images for any irregularities. The highlights of this work are listed down below:

1. The method of feature extraction proposed in this approach uses a parallel computing paradigm to maximise the performance of the computational resources used. It aims to cut down processing time and costs involved in the digital analysis of medical samples.

2. The methods chosen for the description of images in the form of a numerical vector are derived from averaging methods over pixel intensities - ensuring that they are representative of the global features of the image.

3. The results of the analysis through clustering algorithms are easily understood; the utilisation of various visualisations gives a concise yet thorough insight into the precision of the system. Evaluation of the feature extraction method is done in a similar fashion.

Technical contributions aside, this study also provides valuable insights into the histopathology domain. The current procedures and evaluation standards are brought forward, and their limitations are discussed through various referenced works.

1.3 Structure of the Dissertation

The following portion of this dissertation report is divided into four chapters. The next chapter, Chapter 2, briefly discusses the main concepts and current state-of-the-art work which have inspired this research. It further presents several studies that have successfully integrated said concepts in their research in diverse fields. Chapter 3 details the methodology behind the proposed system, discussing the choice and functioning of the several components of the system. The knowledge domain is discussed as well, followed by data acquisition and sources. The architecture design and the technology used to implement the proposed systems are discussed in chapter 4, followed by evaluation of the approach under various scenarios. Chapter 5 concludes this report, outlining the limitations and discussing the scope for future work.

Chapter 2

Literature Review

This chapter elaborates on the motivation behind the work done in this dissertation. It further discusses the related work that has been referenced to gather the concepts used in this dissertation, giving a concise overview of the data that we are working with and the evaluation techniques to be used.

2.1 Motivation

Studies suggest that prostate cancer is the 2^{nd} most life-threatening form of cancer for men after lung cancer [1]. It is estimated that in Ireland, 3,890 males get diagnosed with prostate cancer annually. This indicates that 1 in 7 men in Ireland will develop prostate cancer at some point in their lives [2]. The current screening process varies across various geographic regions, and there is a clear difference in incidence rates in different regions of the world. Some studies [3] attribute it to PSA testing, which is a screening process involving a blood test to check the prostate specific antigen (PSA) levels in a person's blood. Due to extensive PSA testing in the USA and Europe, it is estimated that almost 20-40% cases reported in these regions are due to overdiagnosis [1][4] [5].

Currently, diagnosis involves the use of biomarkers, which are essentially molecules produced by the body of cancer-inflicted individuals. Biomarkers can be of several types, including but not limited to DNA, RNA, protein or metabolic profiles. PSA levels are also a type of biomarker – and a PSA test happens to be the most cost-effective biomarker test available today. However, it is reported that only 3% of all PSA-screened men actually have a lethal condition, leading to an overestimation of the threat prostate cancer poses to men worldwide [6]. Alternative biomarker tests happen to be expensive and time-consuming, leading to delays in the commencement of treatment. They can often be an uneconomical expense incurred by labs, making them an unpopular choice.

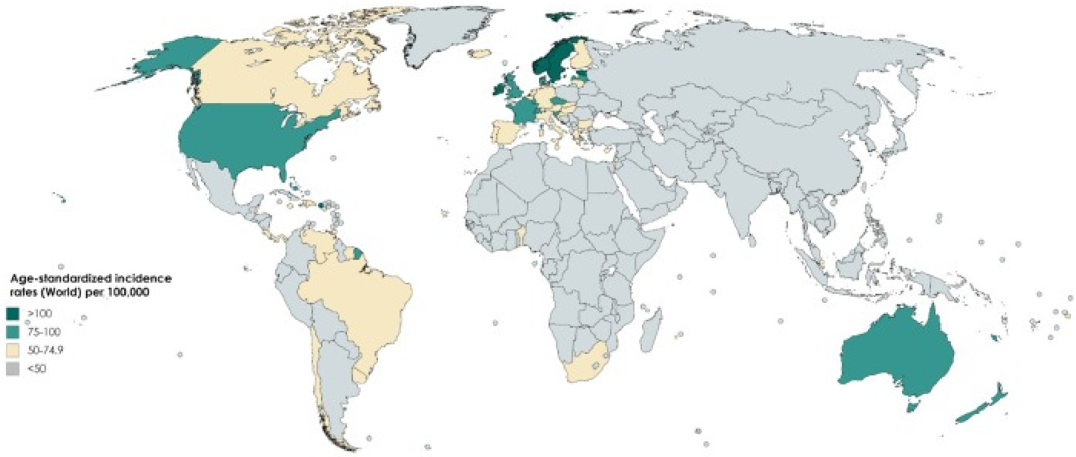


Figure 2.1: Incidence rates of prostate cancer across the world in 2018, as per data collected from Globocan 2018 [1]

In recent years there have been a lot of advancements in the field of histopathology through the use of digital imaging techniques. The digitization of cell and tissue samples currently involves the use of Whole Slide Imaging (WSI) technology. This has potential use in diagnosing and screening several diseases like cancers through analysis of cell and tissue samples taken from the patient. The collection of these digitized samples from patients is relatively simple and inexpensive. However, the major concern with WSI images is that they happen to be very high-resolution images, and as such, they tend to consume a lot of memory for storage, with database sizes going up to several terabytes. Also, the analysis of these images would require focusing on particular aspects or features of the tissue sample, and given the volume of data being generated due to extensive testing, it makes manual examination a very tedious task. This is an issue even with the current standard biomarker tests, which take considerable time when evaluated by human experts. Several works [7][8][9] have proposed ML and image-processing based systems to automate the evaluation aspect of the screening process. But to integrate these systems into current pathological institutions, the volume of medical data to be analyzed must be taken into consideration.

2.2 Literature Review

This section showcases the specifics of studies carried out using various research methodologies, highlighting the major difficulties encountered and their respective benefits, which serve as the inspiration for this work. The literature review has been divided into subsections based on how closely the research and methodologies of the works under examination are related.

2.2.1 Data Management

The analysis of medical images involves the storage and processing of large amounts of data, like high resolution images – the sizes of which may go up to several gigapixels. There are several storage solutions for a database this size and services like AWS S3 and Microsoft Azure Blob are a viable option for this purpose. However, the main issue to be addressed is the processing time that these sets of images will take for any transformation to be applied. A majority of the processing time goes into the feature set extraction from the images, the details of which will be covered later in this report. In this case, a parallel processing framework like MapReduce is a good way to reduce computation overhead on a local system.

The authors in [10] have proposed an innovative MapReduce based framework for the analysis of high resolution WSI images. Taking into consideration the memory requirements of such images, the authors have designed a feature extraction model, aiming to extract the boundaries of the nuclei in each image sample using MapReduce. In the proposed method, rasterized image segments or tiles are first extracted from the complete WSI. These tiles serve as an input to the MapReduce framework, where they undergo image processing, and then aggregation to obtain an output set of combined fragments of the image. The mapper phase of this stage segments the tile to identify different types of objects within the sample, then turns them into boundary vectors or polygons. The reducer phase, using the image ID and object type as the composite key, aggregates the objects of similar types obtained from a single WSI into a single vector. On evaluation, the authors found that the MaReIa's performance scales almost linearly with the size of the dataset.

In [11], the authors have demonstrated the use of MapReduce framework in the analysis of a large dataset of fingerprint images. The objective is to extract a set of features from the images for further use. These features include important identifying attributes of each fingerprint like morphological thinning, ridge ending and bifurcation, each of which is calculated in the mapper function. These features are then aggregated and returned as a set of features by the reducer function. Following this approach, it was observed that the time taken to extract features from a random subset of the fingerprint dataset reduced to half of what it took originally. It is also seen that the processing time per image decreases as the number of images being loaded to the mapper increases, i.e., the performance of MapReduce scaled with the volume of data being processed.

2.2.2 Image Features for Analysis

There are several established methods to convert high resolution pathological images into feature vectors, containing numerical descriptors to better help the classifier distinguish between samples. One of them is SIFT (Scale Invariant Feature Transform), which was proposed by D. G. Lowe in 1999. The SIFT descriptors of an image are basically a local group of features which are scaling, rotation and translation invariant [12]. Additionally, they are robust towards changes in illumination and noise. In this paper the experiments performed use SIFT descriptors to create an object recognition model. Further, these SIFT descriptors are used in another study [13] in combination with Gamma Mixture model to fit a classification model for medical images with texture. In experiments, it was determined that the resulting feature set, called the GCD-GMM features, in combination with SVM or KNN outperformed the state-of-art methods in two datasets [13].

However, it is usually more suitable to take into account the global features of an image when analyzing medical samples so that the gross aspect of the image is preserved [14]. Zernike moments, proposed by Zhang and Lu in [15] are a scaling, rotation and translation invariant feature set derived from orthogonal Zernike polynomials mapped over the unit circle. In [15] the global nature of the features is demonstrated as the computation involves statistically aggregating information from all pixels in the region. Zernike moments have been seen to be more robust towards distortion and noise in studies [16] and have been used in earlier studies related to medical imagery [17][18][19].

2.2.3 Learning methods

The Kohonen Self organizing map, proposed by T. Kohonen [20] is an unsupervised artificial neural network that can be used for clustering unlabeled data. Self-organizing maps operate by creating a projection of a complex set of features (of the input data) into low dimensional, discretized maps. This reduction of dimensionality makes it easier to process the data and reduces computational load, while keeping the topological structure of the data intact.

The work done in [21] describes the design of a content-based image retrieval system for medical images, using SOM as the underlying clustering model to group relevant search results. The dataset used consists of brain MRI images and the features taken into consideration are the color, shape and texture of the images. The query results obtained from this model show a match accuracy of 93.33% on an average. The authors in [22] have focused the development of a Computer Aided Diagnosis (CAD) system, focusing on the diagnosis of breast cancer. They have proposed an alternative to hematoxylin and eosin stained biopsy images in the form of a feature set generated by a pre-trained Convolutional Neural Network (CNN). The resulting feature vector is used to train a Support Vector Machine (SVM) classification model to predict carcinoma. The objective is to classify the images into four categories - normal tissue, benign lesion, in situ carcinoma and invasive carcinoma [22]. Through experiments it was observed that this model was 83.3% accurate in predicting carcinoma/non-carcinoma.

2.2.4 Related Work

In a recent piece of work done by Google Research [23], the authors have developed a content-based image retrieval system that works similar to the google reverse image search, but specifically for medical images. Similar Medical Images Like Yours (SMILY) was developed specifically to medical health professionals in tissue sample analysis. The feature vector of each image is generated through a pre-trained convolutional neural network, which returns a set of numerical values that describe the images. These numerical descriptors are called “embeddings”. The query consists of an image, the embeddings of which are compared with the embeddings of the rest of the images in the database to return the k most relevant results (where k is customizable).

The authors have taken into account that analysis of medical images requires looking at specific regions of the tissue sample. The images acquired can be very large in size and visual analysis of the images requires looking at different aspects of the cell or tissue. Hence the tool provides options to select specific regions of the image to retrieve similar images. Users can also specify the preferred axis of similarity like histologic feature or tumor grade to retrieve images.

Google SMILY is not open source as of now. On reaching out to the authors of the paper, it was found that there are no plans of releasing the software or source code yet.

The authors in [7] have proposed a content-based image retrieval system, by the use of SIFT descriptors as feature vectors for medical images. The proposed system is targeted towards pathologists and it allows them to select specific regions of the sample being analyzed. This is done to provide more control over the granularity of the search. This system is called “sCBIR”, short for content-based sub-image retrieval system and it allows the user to focus more on specific structures in the sample rather than the entire image as a whole. After obtaining the resultant SIFT features, a model based on k-nearest neighbors is used to calculate the similarity between the query image and the images in the database for retrieval. On evaluation it was found that results from manual search and sCBIR had an 80% match on a dataset consisting of 50 stained high-resolution images of prostate tissue.

Structural Similarity Index Metric (SSIM) is a metric proposed by the authors in [8] that assesses the visual impact of three characteristics of an image: luminance, contrast and structure. As such, it works best with grayscale images and doesn't take into account the color channels or hues of the image when computing the index. SSIM index is mainly used for image quality assessment, using another image as reference. The values range between 0-1 - values closer to 1 indicate more similarity to the reference image than values closer to 0. In the original work, the SSIM index was used to compare the relative quality of 344 JPEG and JPEG 2000 images. On evaluation, SSIM outperformed all the other models in a comparative analysis.

This has potential use in the analysis of medical images, as in works like [9],[24],[25] and [26]. In [9] the authors have used the SSIM characteristics of image samples as a feature to analyze a dataset of mammograms and predict the probability of a sample being benign. In [25], a model is proposed to computerize the Hematoxylin and eosin staining (H&E) of pathological tissue samples. This model can also reconstruct a de-stained image from a stained sample. SSIM has been used in this study to evaluate the performance of this model by quantifying the similarity of computer-stained images with manually stained tissue samples. The de-stained images were evaluated in a similar fashion. As with the original paper [8], the work done in [26] also compares the image quality of images at different compression ratios. In this work, the goal is to find out what compression ratio is the most optimal for storage and analysis of whole slide images so that storage and transmission are efficient and image quality is good enough for medical evaluation. SSIM was used to compute the loss in structure and definition in the images at different compression levels.

2.3 Conclusion

The above chapter discusses the literature about the several concepts involved in the development of this thesis, giving an insight into the application of these concepts in various related works. The design aspect of this work is greatly influenced by the papers examined in the above sections. This review has greatly helped in the analysis of the strengths and weaknesses of each of the components of this work, which in turn allowed for a better evaluation of the model.

Chapter 3

Methodology

3.1 Introduction

The era of medical imaging began much earlier than most can people anticipate; it can be said that it started around the 1960s when the first telepathology experiments took place. However, it is in the 90s that significant digitization efforts began, leading to the birth of the term “digital pathology” - indicating the shift of traditional histopathology to a more computerized platform.

At the same time, there have been major developments in the field of image processing and machine learning models focusing on image analysis. The combination of the developments in these two areas has shown immense potential in the development of tools and techniques that aid the field of medicine, and several works have demonstrated this further (as mentioned in chapter 2). From the study of new drugs to its application in the case of long-distance consulting of telepathology, the usefulness of digital pathology in this day and age is quite significant.

Digitization efforts over the past few decades include converting patient health records into electronic health records (EHR), and trends indicate that their prevalence increased by up to 4 folds between 2007 and 2012 [27]. Another major development towards digitization has been the introduction of WSI (Whole-Slide Imaging), which involves scanning glass slides to produce images of samples being studied [28]. This is also known as virtual microscopy, and it is gaining popularity as an alternative to the traditional optical microscope. The WSI technology has made several aspects of medical analysis much more straightforward than traditional microscopy, be it portability, ease of transfer, images retrieval, workload balancing, or sample analysis [28]. Recent studies [25] have even proposed models to automate the Hematoxylin and eosin staining of tissue samples by means of image processing and machine learning models. The idea is to replicate the staining effect on whole slide images using an image mask.

Recent years have also seen a spike in the reported cases of diseases related to cell and tissue abnormalities, particularly cancer. This can be attributed to rigorous testing in certain geographic regions and an increase in awareness amongst the general population. This has inevitably led to the generation of huge volumes of medical samples which have to be stored, processed and analyzed by experts. The current protocol and procedures for diagnosing these diseases are increasingly becoming infeasible as the time and costs involved are huge. The need of the hour is to scale up the medical diagnostics framework to allow for a more efficient and streamlined diagnosis of patients. The progress made in the area of "digital pathology", from WSI samples to cloud storage solutions, is a promising venture in that direction. This allows for a much more detailed analysis of images by quantifying their features and applying state-of-the-art algorithms to extract further information, something which was never possible with traditional microscopic imaging.

Figure 3.1 briefly describes the operations involved in the pipeline of the proposed model.

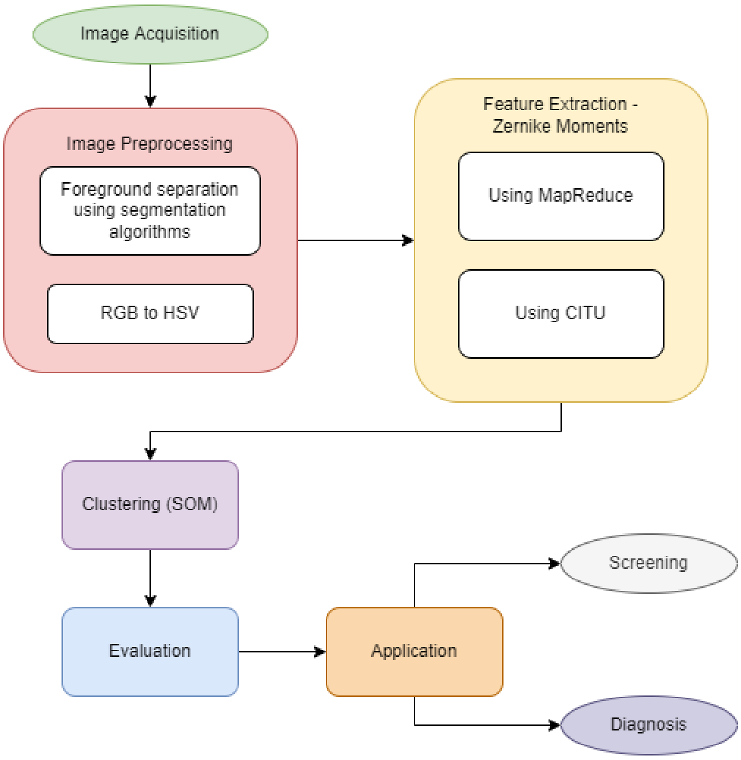


Figure 3.1: Workflow Diagram of Proposed System

The rest of the chapter is organized as follows: Section 3.2 explains the dataset features and steps involving dataset acquisition, while the next section (section 3.3) elaborates on the image pre-processing done on the dataset to enhance its feature. The transformation of image dataset into a quantifiable feature vector is explained in the next section (section 3.4) followed by the methods utilized to extract these features from the images (section 3.5). Finally, the clustering approach used in this work is discussed (section 3.6) along with evaluation (section 3.7) and conclusion (section 3.8) of this chapter.

3.2 Domain Knowledge and Data Acquisition

The first step toward analyzing data is gathering knowledge about the dataset itself. Without sufficient information about the domain, it is impossible to determine what features have to be extracted from the data to be passed as an input to the clustering model. It is equally important to know the data source and understand the conditions under which the samples were taken. This would help identify any irregularities or outliers in the dataset and, if possible, find methods to eliminate them. The evaluation of the system performance is also hugely dependent on these aspects of the research.

The domain that we are working with in this study is medical image analysis – to be specific, it's the analysis of prostate tissues. As such, the dataset that we are working with comprises entirely of high-resolution WSI images of Hematoxylin and Eosin stained prostate tissue. In a previous study [29] that used a similarly-sourced dataset as this work, interviews were conducted with experts to understand the dataset at hand and the challenges faced by human pathology experts in evaluating this kind of data. This subsection discusses the major highlights of these interviews and in the process, supplements the understanding of the domain that is the focus of this research.

As per [29], there are certain factors taken into consideration during the histopathological analysis of prostate tissues, some of which are outlined below:

1. The principal stain used in the obtained tissue samples is the Hematoxylin and Eosin stain, also called H&e stain. Staining of samples is usually done to highlight certain features of the sample like nuclei and cytoplasm. H&e staining is very common in the analysis of tumour tissue microscopic analysis – the hematoxylin dyes the nuclei blue by binding with the DNA, and eosin dyes other parts of the tissue like stroma and cytoplasm pink, in addition to turning the red blood cells dark red [30]. This facilitates better visual analysis of samples by enhancing the distinguishing features

of the sample. In the case of cancer tissues, this stain gives a bluish tint to secretions in the lumen and sometimes pink crystalline structures are observed.

2. From a pathologist's point of view, that analysis of H&e stained prostate tissue involves looking at certain structural features of the WSI. These include – small glands in front of large glands, haphazard distribution, high density of nuclei and/or large nuclei and presence of prominent nucleoli.

3. H&e tissue analysis requires experts to examine samples under 40x-400x zoom depending on the complexity of the features.

4. One of the methods of evaluating the severity of a diagnosed prostate cancer case is the Gleason grading system. Gleason scores are assigned based on the histological patterns observed in the gland tissue structure. The scores can range from 1 to 5 based on the type of growth pattern observed in the tumour biopsy, 1 being the most favourable prognosis indicating malignancy and 5 being the least favourable indicating aggressive cancer. The total score is the sum of the scores obtained through the grading of the two most predominant patterns observed in the sample – e.g. a score of 3+4 = 7 indicates an overall Gleason grade of 7, 3 being the score of the primary pattern and 4 being the score assigned to the secondary pattern. If there is no secondary pattern, the score of the primary pattern is added up twice, for e.g. 3+3 = 6, where 3 is the score of the primary pattern.

3.2.1 Data Acquisition

The dataset used in this study was directly provided by Dr Aamir Ahmed, Head of the Stem Cell and Prostate Cancer Group at King's College London. The primary research interests of Dr Aamir lie in the Wnt signalling network and the role it plays in prostate cancer and prostate stem cells. Through his group's research, a new class of drugs termed membrane potential regulating compounds (MPRCs) have been identified, which are currently in use for the treatment of diseases other than cancer and have potential use in the treatment of prostate cancer. For research purposes, the group has used high resolution, high throughput tissue imaging in combination with a gene, single-molecule RNA and protein immunochemistry, and machine learning for the identification of biomarkers that aid the diagnosis of prostate cancer[31][32][33].

The dataset consists of 111 high-resolution Whole Slide Images of prostate tissue, collected anonymously for research purposes. Slides were magnified at 40x with the

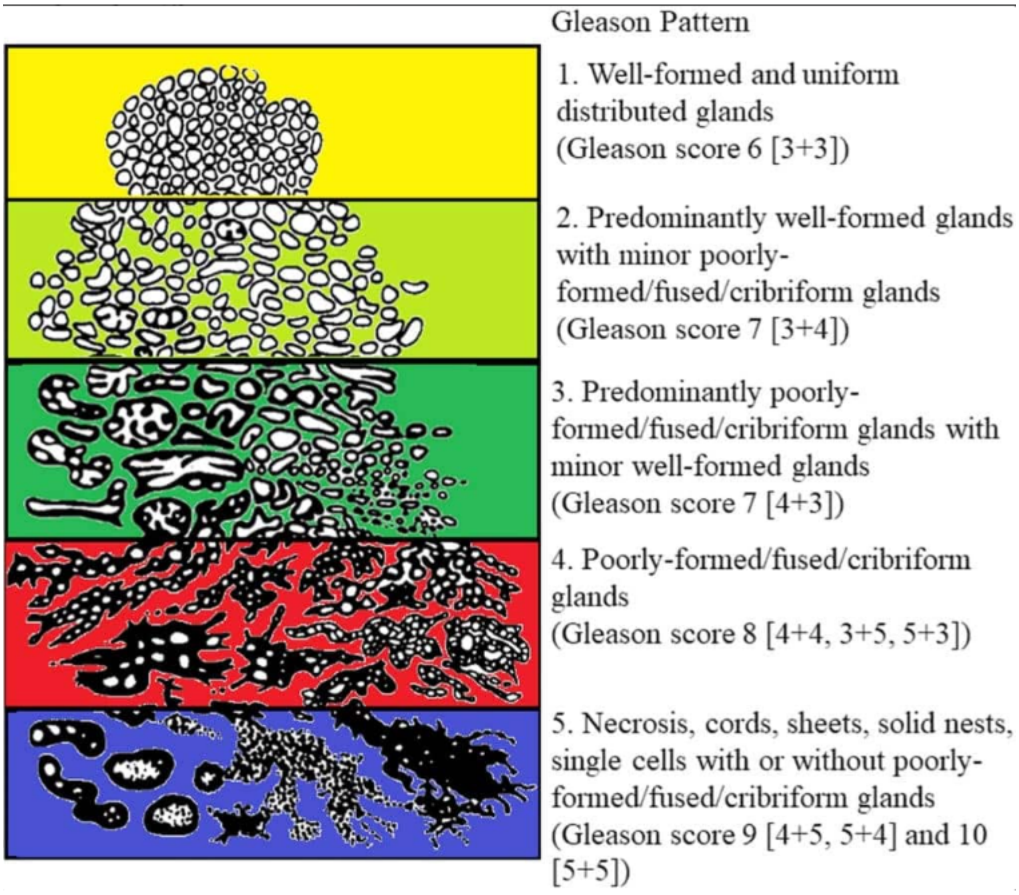

Gleason Pattern
1. Well-formed and uniform distributed glands (Gleason score 6 [3+3])
2. Predominantly well-formed glands with minor poorly-formed/fused/cribriform glands (Gleason score 7 [3+4])
3. Predominantly poorly-formed/fused/cribriform glands with minor well-formed glands (Gleason score 7 [4+3])
4. Poorly-formed/fused/cribriform glands (Gleason score 8 [4+4, 3+5, 5+3])
5. Necrosis, cords, sheets, solid nests, single cells with or without poorly-formed/fused/cribriform glands (Gleason score 9 [4+5, 5+4] and 10 [5+5])

Figure 3.2: The Gleason Scale indicating different grades of cancer corresponding to different tissue structures

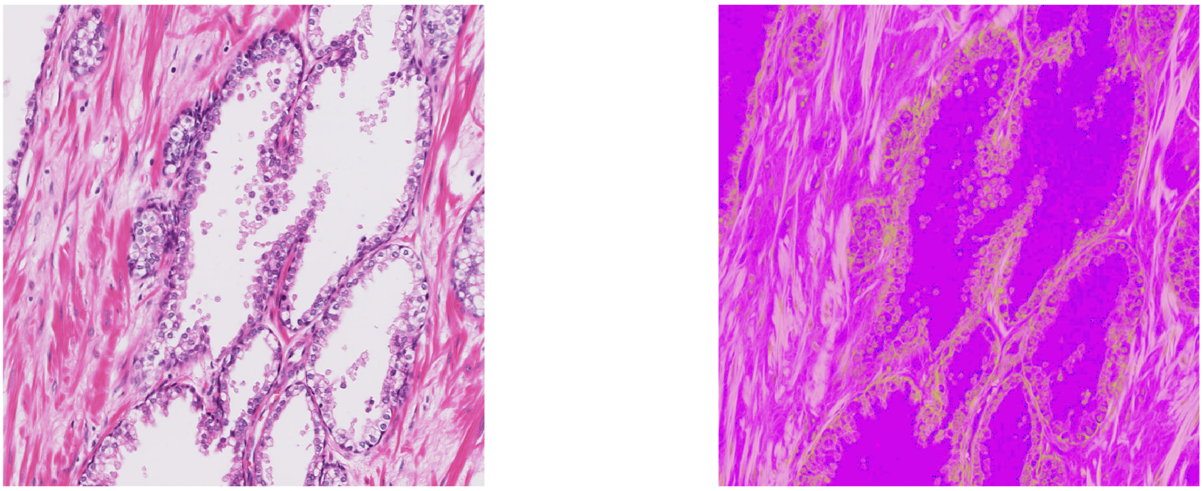
Nanozoomer slide scanner (Hammamatsu Photonics UK Ltd, Welwyn Garden City, UK), a high-resolution scanner, to produce these images and digitize them. This kind of technology allows for extremely high-resolution images giving a digital magnification of up to 60x, contributing to the large file size of the images. The tissue samples used had undergone h&e staining before the scan to bring out the distinguishing features of the samples.

3.3 Image Pre-processing

Many studies dealing with the analysis of images find it useful to apply pre-processing of some kind to them before examination. This is done in order to either denoise the image, correct distortions, preserve structural or colour information or enhance certain features contained within the image.

In this study as well, the application of pre-processing techniques has the potential to greatly improve the overall performance of the proposed clustering system. The images in the prostate tissue dataset are high-resolution images containing several uniform or non-uniform structural patterns. A lot of the features of these images are also described by the colour information present in them. If this dataset is directly used as an input to a clustering model, there is bound to be some loss of information during the vectorization process, and the model may perform poorly because certain distinguishing attributes were eliminated from the input.

To counter these abnormalities, this study could benefit from certain image pre-processing techniques, one of them being the RGB to HSV colour channel transformation. Several studies have indicated that the HSV encoding is much better in preserving information intrinsic to the colour characteristics of an image [34]. This kind of augmentation also helps provide better contrast and colour representation. Figure 3.3a and figure 3.3b highlight the visual differences after changing channel encoding.



(a) RGB encoded image sample

(b) HSV encoded image sample

Figure 3.3: RGB vs HSV encoded prostate tissue samples

On observing the image samples, it is evident that there is a clear distinction between the foreground and the background. The objective of the study is to analyze prostate tissue, which constitutes the foreground of the image, meaning that that background has no contribution whatsoever to the feature vector. Naturally, it makes sense to remove the background pixels altogether, which in turn will reduce image size and cut down processing time. This can be achieved through several methods, one of them being image segmentation. There are several ways of segmenting images, like thresholding, edge-based segmentation, region-based segmentation or clustering-based segmentation.

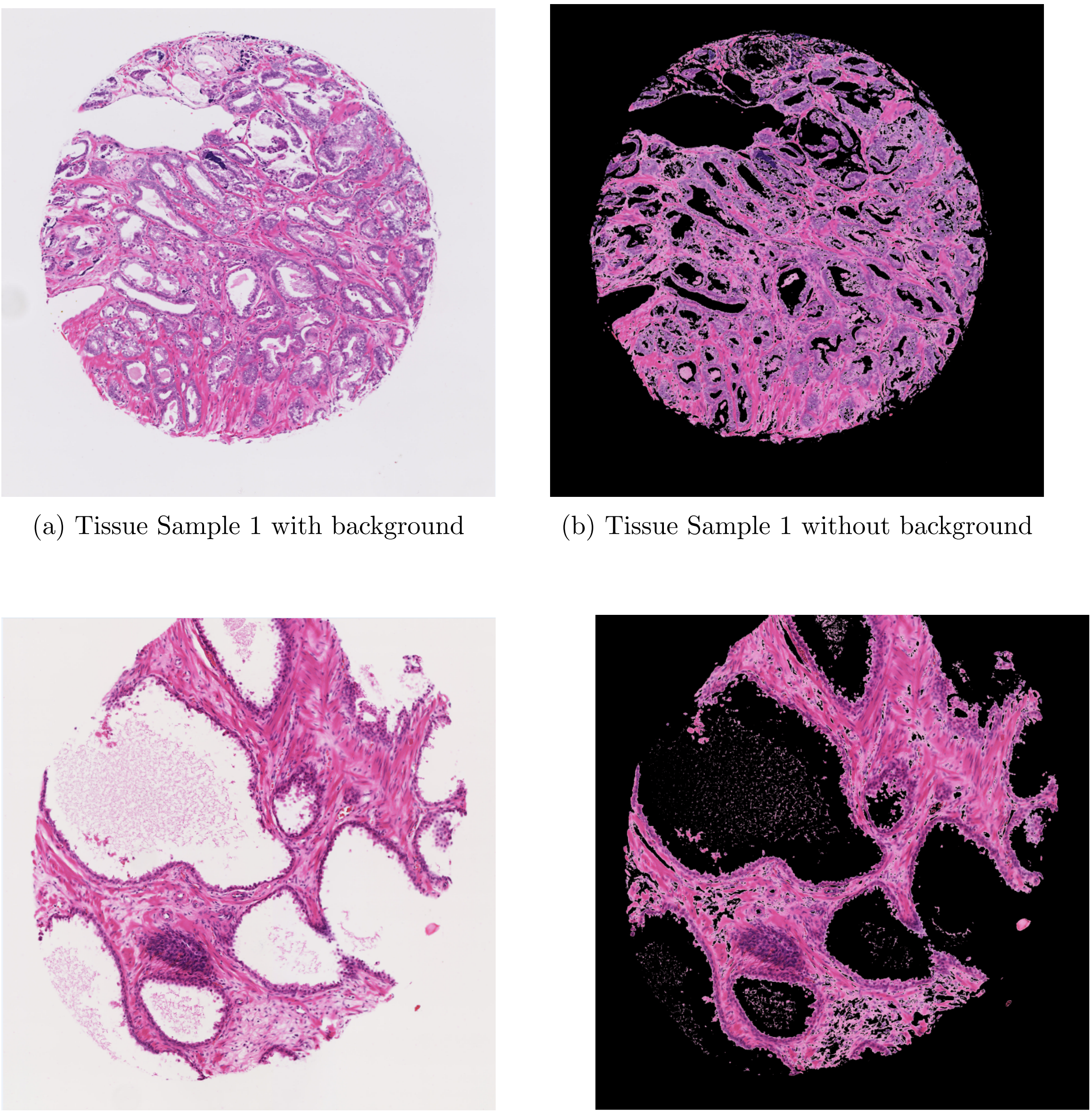
Studies have been done that indicate the inefficacy of thresholding and edge-based segmentation methods, in that they are unable to properly segment images with irregular edges [35]. Since we have image samples with tissue structures ranging from what can be described as well-formed uniform (healthy tissues) to very distorted and poorly formed (cancerous tissues), these two methods are not suitable for this purpose. Thresholding also doesn't guarantee that the segmented regions are contiguous [35], and for the purposes of this study, this is not ideal. Region-based approaches happen to be computationally expensive and time-consuming and require the manual selection of seed regions based on which regions are determined. This is inefficient as it does not fall in line with this work's objective of automating the entire analysis process. This leaves clustering-based methods, which is what has been used in this study.

K-means clustering is a popular method for segmenting colour images. It is a relatively simple algorithm with a low computational overhead, and the segments produced by this algorithm have no overlap. K-means clustering is an unsupervised clustering algorithm. In its application in image segmentation, it is able to cluster similar pixels together, making it ideal for the separation of foreground from background in the images used in this work. Other authors have cited the use of this algorithm for image segmentation in medical images owing to the fact that the refinement of the segmentation can be customized by specifying the number of clusters prior to execution [36] [37].

In this study, however, segmentation using k-means clustering requires only two non-overlapping segments of the image – namely, the foreground and the background. Thus, only two clusters are specified. Figure 3.4 shows the results of this segmentation process. These images indicate clear segmentation between foreground and background, and the background pixels are completely removed after this step. The segmentation results are accurate irrespective of the shape or structure of the tissue.

3.4 Image Representation as Feature Vector

Post the enhancement of image features using pre-processing techniques, there is a need to convert these images into numerical feature vectors. This is done in order to be fed as an input into the clustering model, as raw images cannot be read directly by these systems. There are several ways to quantify image features and combine them to form a vector – a rudimentary approach to do so would be by looking at descriptive attributes like pixel intensities, luminance, shape, texture, etc. However, these attributes are not true descriptors of any image as they do not capture all the details contained within



(a) Tissue Sample 1 with background

(b) Tissue Sample 1 without background

(c) Tissue Sample 2 with background

(d) Tissue Sample 2 without background

Figure 3.4: Comparision of tissue sample before and after foreground seperation

it. Methods like wavelet transform are well accepted in the analysis of images [38][39]; however, this method has a known drawback for performing poorly in representing and detecting object contours in images. All the traditional methods mentioned above do not scale well with large datasets, which is a major concern that our study is dealing with. Due to the above reasons, this work has looked into different methods of representing images numerically. Of the several alternate techniques explored, the concept of image moments stands out as the characteristics of these features fall in line with the objectives of this research. The concepts behind these moments have been elaborated on, and advantages are highlighted further in the below sub-sections.

3.4.1 Image Moments

Moments are a common concept used in several fields like statistics, mechanics and many more. Moments are mathematical representations that describe the characteristics of any distribution of points. In image processing, moments are a set of specific weighted averages of the pixel intensities of the image. First introduced by Hu in 1962 [40], moments have found their use in image analysis in various studies for the purposes of pattern matching, object recognition, etc. [41]. Each moment can describe a particular characteristic of the image, like area, centroid, orientation, etc. Since moments are derived from the weighted averages of pixel intensities, they are able to capture the global information contained within each sample. Image moments can be further utilized to derive moment invariants, which are essentially invariants with respect to certain transformations like scaling, rotation and translation.

The raw moment m_{pq} of order $(p+q)$ for a continuous function $f(x,y)$ is defined as

$$m_{pq} = \int_{\infty}^{-\infty} \int_{\infty}^{-\infty} x^p \ y^q \ f(x,y) \ dx \ dy \tag{3.1}$$

For a grayscale image with pixel intensities I(x,y), the above formula translates to

$$m_{pq} = \sum_{x} \sum_{y} x^p \ y^q \ f(x,y) \tag{3.2}$$

3.4.2 Moment Invariants and Orthogonality

To understand the benefits of the use of moments to represent images in the form of moments, we must first understand what moment invariants are. A moment invariant in image processing is a characteristic of the image that does not change or only slightly changes if the image is transformed (e.g., rotated, scaled, blurred, etc.). This means that the moments derived for two images of the same object with different scales, rotation degrees, or geometric shifts in position will be the same. This property is very useful in this study as the images in the used dataset are prone to slight aberrations due to differences in the positioning of the foreground and background and being captured at different angles.

Figure 3.5 shows 3 images of the letter S and one of letter K. By the principles of moment invariants, one will find that there is no difference in the moments derived from the first three images showing the letter S, even though there is a difference in the rotation, scaling and geometric positioning of the subject. However the moments for the letter K will differ from that of the letter S in the three images as the subject has changed.

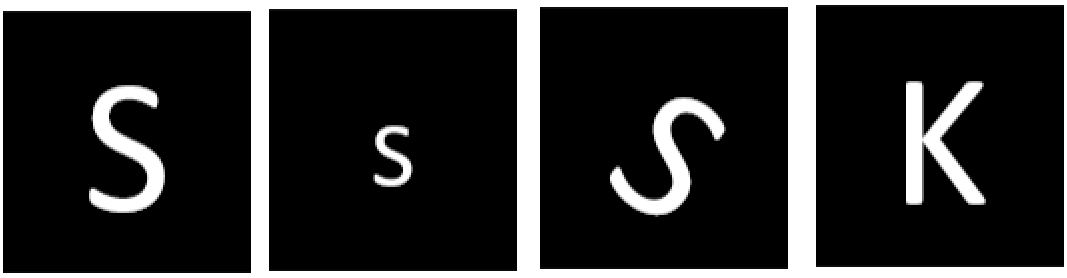


Figure 3.5: Three images showing the letter S at different degrees of rotation and a single image of letter K

Another property that aids in the accurate analysis of images is orthogonality. Orthogonality in the context of image moments can be described as the property of moments being uncorrelated to each other. Due to this property, each moment is independent of the other and there is no overlap of information represented by a set of orthogonal moments. Regular image moments are not orthogonal and contain a lot of redundant information, which makes calculation computationally expensive and tedious.

3.4.3 Zernike Moments

Zernike moments were first proposed by Zhang and Lu in 2004 [15]. Zernike moments of an image are derived by Zernike polynomials which are a set of polynomials that are orthogonal to each other on a unit disk. Zernike moments are a preferred way of image representation in the field of image analysis as they can be mathematically represented just like regular moments but have the added benefit of being orthogonal, This makes for more efficient computation and accurate representation of the image.

Zernike polynomials are calculated differently for even and odd counts. Below is the formula for calculating even Zernike Polynomials over a unit disk:

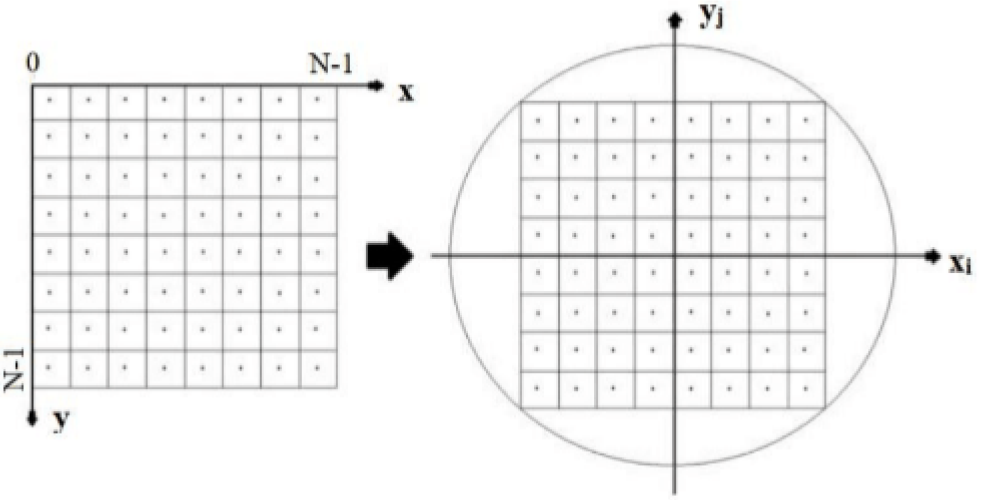


Figure 3.6: An image that is completely mapped inside the unit circular disk is referred to as outer circle mapping

$$Z_n^m(\rho, \varphi) = R_n^m(\rho)\ cos(m\varphi) \tag{3.3}$$

And below is the formula for calculating odd Zernike Polynomials over a unit disk:

$$Z_n^{-m}(\rho, \varphi) = R_n^m(\rho)\ sin(m\varphi) \tag{3.4}$$

Where,

n is the order of the Zernike Polynomial,

m is the repetition,

m & n are non-negative integers where $n \geq m \geq 0$,

ρ is the radial distance of point from the center of the image,

φ is the azimuthal angle between point and x-axis, And R_m^n is the radial polynomial defined below:

$$R_n^m(\rho) = \sum_{k=0}^{\frac{(n-m)}{2}} \frac{(-1)^k\ (n-k)!}{k!\ (\frac{n+m}{2}-k)!\ (\frac{n-m}{2}-k)!} \tag{3.5}$$

Zernike moments of different orders and repetitions describe different properties of the image. Some are listed in table 3.1.

For a particular image, the Zernike moment is computed by projecting the pixel coordinates to the unit circle's range, where the origin of the unit circle is the center of the

Table 3.1: Zernike moments and the different properties of an image they describe

Zernike Moment	Property
z00	Piston or Area
z11	Horizontal tilt
z02	Defocus
z13	Vertical Astigmatism
z22	Horizontal Coma
z24	Vertical secondary astigmatism

image. As such, the Zernike moments are known to not capture information about pixels that fall outside the unit circle.

Digital pathology involves the examination of images that contain a variety of visual information - details consist of distinctive features like shape, tissue structure, patterns, etc. work done by Zhang and Lu has shown that Zernike moments are successful in capturing most of this information. The calculation is done by averaging of pixel intensities, hence these features are representative of global features of the image and are not focussed on local aspects only. Orthogonality of the features allow for computational efficiency as redundancy is eliminated. Zernike moments also have the added benefit of noise elimination in images as the computation is based on a summation process.

This study has only used lower order Zernike moments, upto the order 12, due to the fact that higher order Zernike moments are susceptible to capturing noise in the image. There are several studies that have corroborated the same, be it facial recognition systems, content-based retrieval systems [42] [43] or pattern matching systems [44]. The authors in [44] used two different datasets - one with noise and the other without, to find out the optimal Zernike descriptors for the data in question. The work done in [42] and [42] have attempted the same by observing the performance of a content-based retrieval system. Each of these studies concludes that lower order Zernike moments prove to be better descriptors for the images in question and are less likely to capture noise in calculations. Also, there are certain Zernike polynomials that are not allowed as the condition $n \geq m \geq 0$ should always hold true.

Overall, Zernike moments have been effectively employed in several applications and show an advantage in capturing texture and form aspects of these medical pictures. This is owing to its many advantages, including compact features, excellent accuracy, cheap computing cost, and robust feature retrieval.

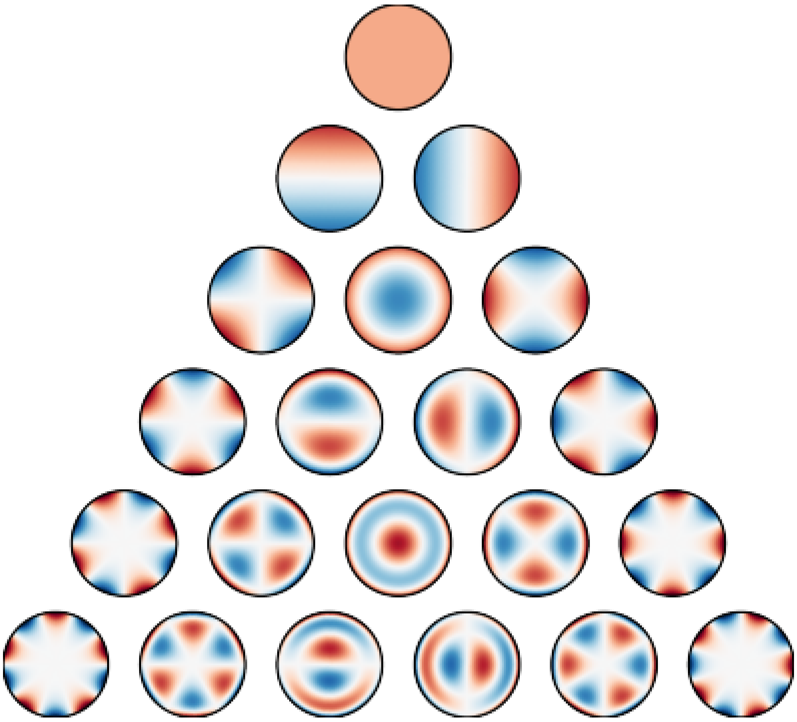


Figure 3.7: The first 21 Zernike polynomials, ordered vertically by radial degree and horizontally by azimuthal degree

3.5 Feature Extraction Methods

As mentioned previously, the major motivation behind this study is dealing with the huge volumes of medical data generated due to extensive testing. Digitization efforts in pathology have led to a huge surge in the generation of high resolution medical images, all which must be stored, cataloged and processed, either by automated systems or by human experts. The latter seems to be more prevalent in current times as there is very less room for error in any analysis conducted on medical data. One way to cut down on the cost and time consumed in this endeavor is to utilize powerful parallel processing frameworks to help extract meaningful features from the image, which can be further utilized for analysis by professionals or computerized systems. One such framework, called MapReduce, has been gaining popularity in this field and there have been studies that have implemented this framework for feature extraction and analysis [6][25].

3.5.1 MapReduce

MapReduce is a parallel processing framework which is mainly used to work with large datasets that do not fit in-memory. It was popularized by Apache Hadoop as an important tool in the Hadoop Framework for the processing of data within the Hadoop Distributed Framework System (HDFS).

MapReduce framework is generally used for computational problems that need parallelization of operations on datasets that are spread out over many nodes, generally referred to as a cluster if all the nodes operate within the same local network and have similar hardware. It can also be used when the data in question is distributed across a grid, which is essentially a group or nodes in geographically disparate locations having distinct hardware specifications. This study follows a slightly different implementation of MapReduce, which is available in MATLAB as the mapreduce function.

The first stage in processing data with the use of MapReduce involves the construction of a datastore, which is used to individually work on small chunks of the entire dataset that fit in-memory. There are two important phases in the MapReduce programming paradigm - namely Map and Reduce. The Map phase involves getting the input data, applying transformations and storing these intermediate results in a temporary storage. The Reduce phase is primarily used to aggregate these results and extract meaningful information from them, and store them as the final output.

3.5.2 MapReduce Architecture

Figure 3.8 shows the overall architecture of the MapReduce framework architecture. As mentioned before, the input datastore is responsible for the storage of the entire dataset that is being processed, in the form of several chunks, each of which can be held in-memory. Each chunk of data is passed on to a mapper, which performs some sort of precursory computations on it and passes it on to an intermediate storage object. There are many mappers at work in parallel, and the number of mappers is equivalent to the number of chunks of data in the input datastore. The mappers store data in the intermediate datastore as key-value pairs, where each key can be linked with many values. This is done for the purposes of grouping similar data together in the next stage, based on problem specifications.

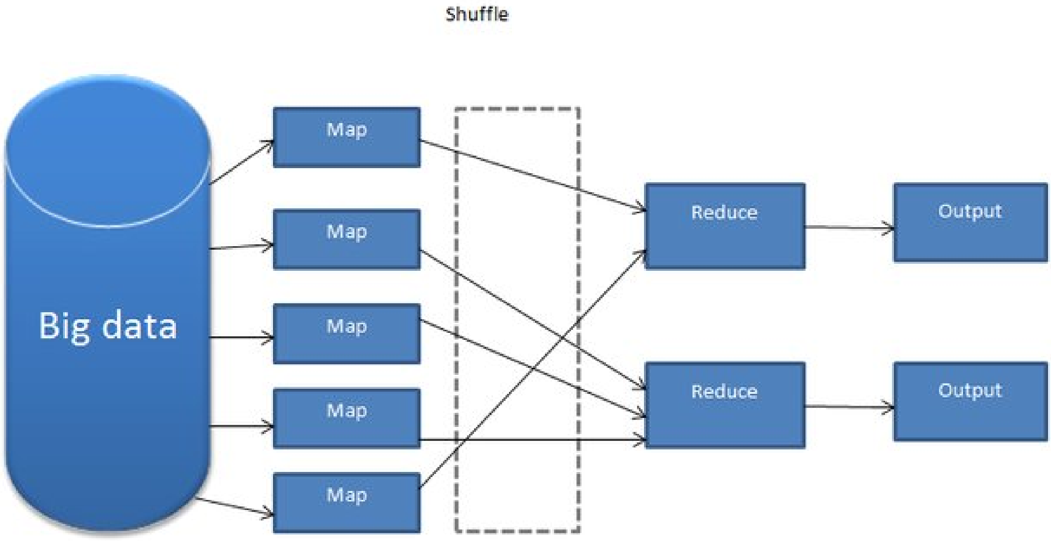


Figure 3.8: MapReduce Framework Architecture Diagram

This intermediate data then goes through the shuffling and sorting phase, which is where values with the same key are grouped together. This stage consumes the output of the mapper phase and its results serve as an input to the reducer phase. It is here that the output blocks from the mapper designated with specific keys are assigned to each reducer.

In the reduce phase, the values associated with each key are aggregated and meaningful information extracted. Aggregation could mean any statistical information derived from a group of data points, like mean or sum or even more complex operations. It could also be as simple as just capturing the values linked with a key in a list or vector. The reducer phase also involves the formatting of output data as per requirement.

3.5.3 Proposed Model

For the purposes of this study, a MapReduce based feature extraction model has been proposed, which it briefly described in Figure 3.10.

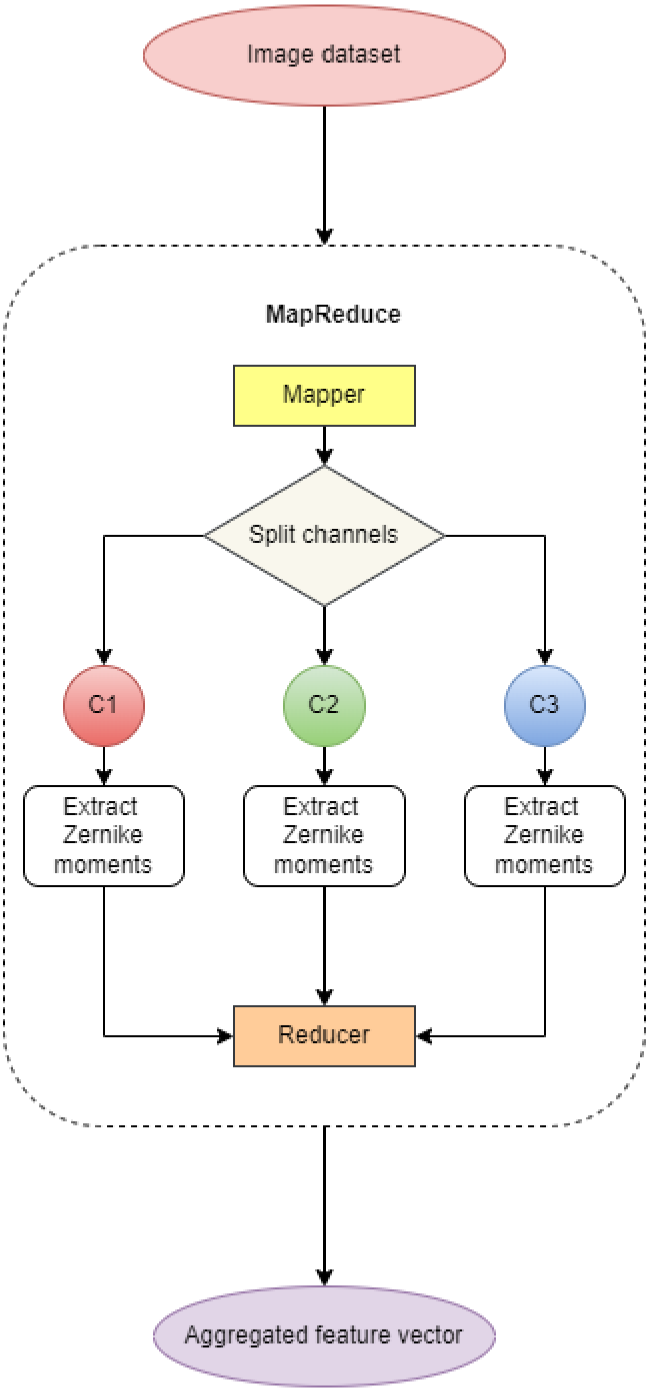


Figure 3.9: Proposed feature extraction model using MapReduce framework

The proposed model aims to reduce the complexity and resources involved in the image feature extraction stage by parallelizing the process across the 3 channels of the image (RGB or HSV, depending on the pre-processing done on the image). The images are stored in the input datastore and passed on to the mappers directly. In the map phase, the image pixels are segregated into the different channels it is encoded in and zernike features extracted using pre-defined computations. Each feature vector for each channel is then added to the intermediate storage object, with channel names as keys. These keys are then used in the reducers to aggregate feature vectors for each image across all channels.

This architecture uses the principles of MapReduce to divide the feature extraction task into 3 separate tasks and execute them in parallel. This method of feature extraction allows the system to extrapolate information from larger datasets in a more efficient manner as only small chunks of data are processed at a time, and the entire data does not need to be at the same physical location for the entire process.

3.5.4 Performance Considerations

A key consideration when using MapReduce to reduce computational overhead is the size of the dataset being processed. Most of the time, utilizing MapReduce paradigm is ineffective for tasks that can be completed quickly and where the data can fit in the main memory of a single machine or small cluster. This is due to the complexity of the operations that enable parallelization in this framework.

Consider an example of a simple task like finding the count of each word in a document. With a traditional iterative approach, the implementation would require running a loop over the entire document and simply increasing the counter for each word as it is encountered. MapReduce framework however would require the instantiation of a datastore, several mappers and reducers. The operations involved require the linking of chunks of data to keys, shuffling the chunks and then performing aggregation operations. On the surface this looks like a more computationally expensive process that consumes memory and requires unnecessary intricate coding patterns. However, it is to be expected that the execution time for the iterative approach scales linearly with the size of the corpus, while with MapReduce, division of tasks between several mappers and reducers ensures that the computational overhead is low with respect to the size of the dataset. When bigger volumes of input data is processed, the memory consumed by the instantiation of several different components in the mapreduce framework is not as big a concern as the memory used by the operations involved.

As such, this framework is deemed suitable for the analysis of medical images in the aforementioned studies and can prove to be useful in this research as well.

3.6 Clustering Methods and Evaluation

The diagnosis of prostate cancer is aided not only by finding the optimal methods to store and process data, but also by figuring out ways to automate the diagnosis - computerizing the evaluation of data contained in each sample. Several attributes of an image sample are of key importance here; analysis from the perspective of an expert pathologist requires looking at structural patterns, inconsistencies in color and design and so forth. Human evaluation of these images is limited by several factors like noise present, quality of the image, complexity of the disease and most importantly, the number of images that can be processed by an individual before the point of exhaustion. There have been many developments in the field of Computer Aided Diagnosis (CAD) and developed systems have since been used commercially for this very purpose, however, no system is an exact fit for all use cases. Examination of different datasets taken under different lab conditions require different CAD systems designed and tuned to handle these inconsistencies. Once the evaluation is completed by any of these systems, the results are examined by human experts for the final diagnosis.

Given the progress made in the field of Machine Learning based image processing , it makes sense to look into Machine Learning based systems to achieve similar objectives. The approach used by ML based systems is to check given data and look for similar features or patterns, then utilize this similarity information to predict results for unencountered use cases and assist in decision making. One of the proposed models that could prove to be useful in the case of diagnosis based on medical imaging is the Artificial Neural Network (ANN). Artificial Neural Networks (ANNs), a subset of machine learning, have demonstrated abilities in pattern recognition that may be used to assess such medical pictures. They mimic the neural connections in the human cognitive system to make decisions. These systems, which were modeled after the real brain, are excellent at learning new tasks by taking into account example data without being specifically programmed to do so.

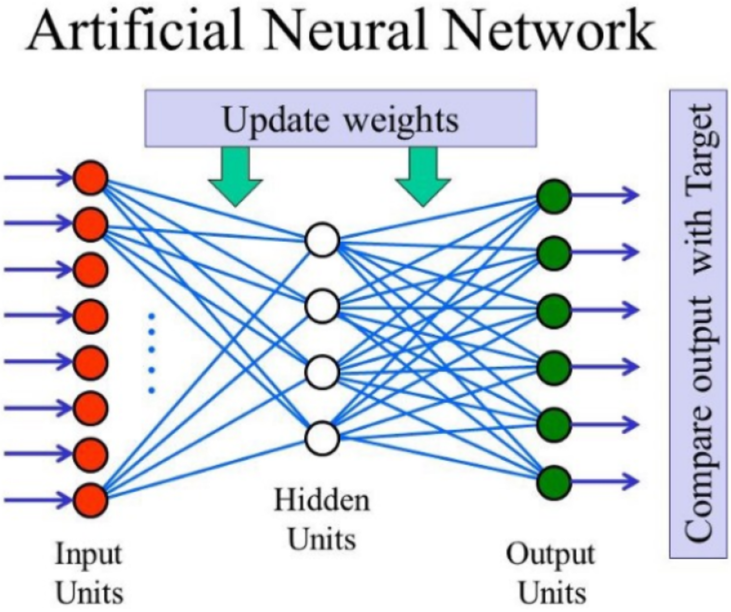


Figure 3.10: Artificial neural network architecture

Much of the theory and conceptualization that went into the development of ANNs was done decades ago, but it is only now that the full potential of this architecture is being realized. This is due to the fact that computational resources like GPUs are much more accessible today and huge volumes of data are made available, which is crucial for better training of ANN based models.

The learning or training process of these networks rely on the training examples they are fed, each of which contains a known input (feature vector) and output. With these examples the network learns to form probabilistic associations between the two, and using this learned information tries to distinguish between any new inputs it encounters. The learning process involves calculating the difference between the predicted output and the actual one, which is known as the error. Based on this error the network is supposed to make adjustments in the weights corresponding to each of the associations formed.

ANNs form the basis of the clustering approach called Self Organizing Maps (SOM) which will be discussed further in the following sections.

3.6.1 Learning Methods

To develop an accurate model for the evaluation and screening of prostate cancer, it is imperative that the clustering approach chosen is compatible with the feature vector extracted from images. The choice of clustering approach to be used for the objectives of this study is dependent on the type of dataset available. Machine learning systems are broadly classified into supervised and unsupervised algorithms, and sometimes a hybrid approach called semi-supervised clustering is also used. The main difference between the three approaches is the labeling of training data - supervised methods involve the use of labeled data, unsupervised methods are used to work with unlabelled data and semi-supervised methods are used when only a portion of the data is labeled while the rest is unlabeled. Supervised learning is aided by the presence of labels; the evaluation of supervised systems is easier as a comparative analysis can be done by comparing the actual and predicted outputs side by side. Supervised learning is also supplemented by the presence of output labels as the error can be predicted during training and be improved upon. In contrast, unsupervised learning methods do not predict on the basis of labels, instead they try to group similar items together on the basis of certain parameters like Euclidean or Manhattan distance between feature vectors. As such, it is difficult to do any objective evaluation on the predictions, except analyzing the items in different clusters.

As the dataset we are working with has no labels, i.e, we don't know which tissue sample is healthy and which is unhealthy, unsupervised learning methods are used in this study. This has the added benefit of being convenient for pathologists analyzing large datasets as very less manual intervention is required to train and test such a model.

Clustering, the unsupervised learning approach used to recognise patterns in the data and group together similar items in same clusters is suitable for the analysis of the prostate tissue dataset at hand. There are several well researched methods for clustering this type of data, some of the popular ones happen to be Self Organizing Maps, K-means Clustering and Hierarchical Clustering. Self Organizing maps have been used in several studies involving the analysis of medical images as mentioned in section 3.2.3. K-means clustering a simpler algorithm which is more efficient in doing color based clustering, as in the case of image segmentation described in section 3.3.

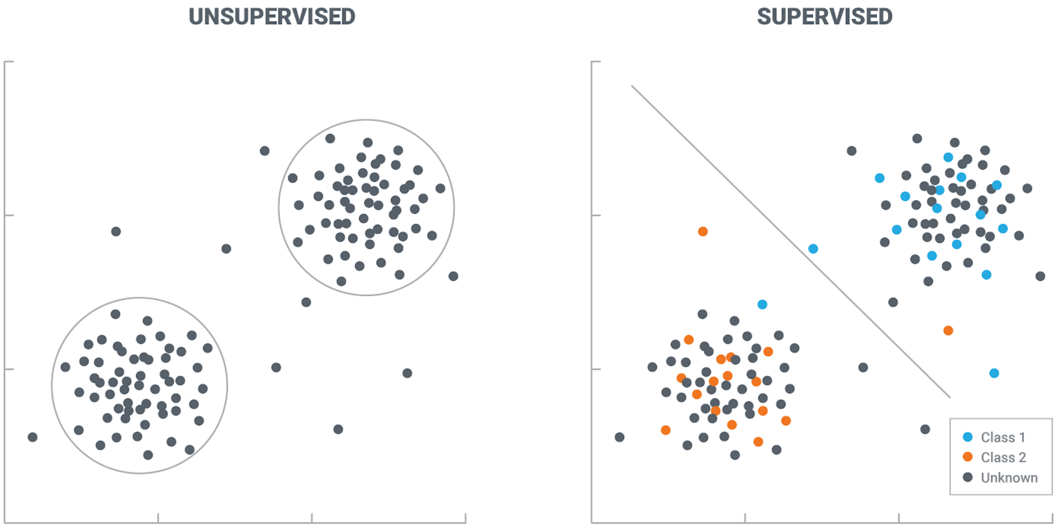


Figure 3.11: Types of Machine Learning based Clustering Approaches, Visualized

3.6.2 Self-Organizing Maps

A self organizing map is an unsupervised artificial neural network that can be used for clustering unlabelled data. Self organizing maps operate by creating a projection of a complex set of features (of the input data) into low dimensional, discretized maps. This reduction of dimensionality makes it easier to process the data and reduces computational load, while keeping the topological structure of the data intact.

SOMs are essentially an Artificial Neural Network with two main layers - the input and output layer. The input layer consists of vectors representing the data points in the input dataset. There are weight vectors associated with each node in the output layer, and the weight vectors have dimensions equivalent to the input vectors.

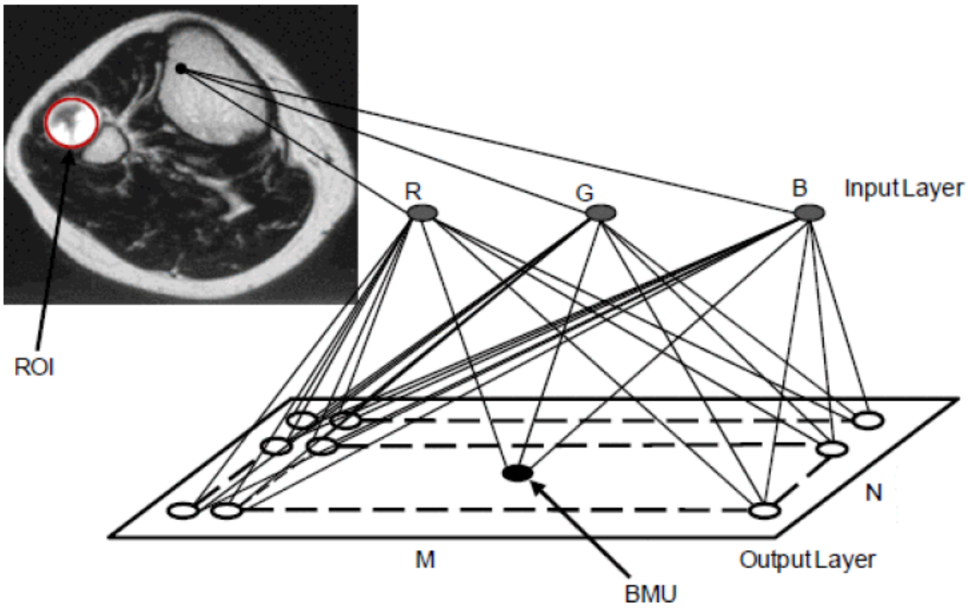


Figure 3.12: Projection of input layer to output layer by SOM

Fig xyz illustrates how the self organizing map projects the features of an RGB color coded image onto the output layer of dimensions $M * N$. This is just an example where the color attributes of the image are used as an input vector, any other set of descriptors could have been used instead.

To begin, let us assume that $W_i = [w_{i1}; w_{i2}; w_{i3}]^T$ denotes the weights associated with the output layer, where $0 \leq i \leq M \leq N$. Initially, these weights are random values assigned at the beginning of execution. The objective of the SOM is to find the Best Matching Unit (BMU) in every iteration, the BMU being the neuron having the values as closest to the input vector. Additionally, the weights of the neighbor nodes are adjusted to increase the likeness with the BMU, bringing them closer to the input vector.

For a node v, the weight update formula is given by

$$W_v(s+1) = W_v(s) + \theta(u, v, s).\alpha(s).(D(t) = W_v(s)) \qquad (3.6)$$

where,
s is the current iteration,
t is the training sample,
u is the BMU for the input vector,
$D(t)$ is the input vector corresponding to training sample t,
$\alpha(s)$ is the learning coefficient,

$\theta(u, v, s)$ is a gaussian function which returns the distance between the neurons u and v in iteration s, also known as the neighborhood function

A brief overview of the steps involved in SOM's clustering process is given in below points:

1. Weights are randomly assigned to the map's nodes
2. A single sample is chosen a current input vector
3. The classification process in SOMs is done by 'competition', which means that the output node with the weights closest to the input vector is chosen as the winner amongst all output nodes.
4. The winner node is known as Best Matching Unit (BMU).
5. Based on this winner node, weight vectors of neighboring nodes are also updated, with the ones more similar to the winning node being rewarded and the ones that are not similar being penalized, i.e, greater amount of weight adjustment is done on nodes that are closer to the BMU and the lesser weight adjustment is done for nodes far away.
6. The above steps ensure that the number of neighbors for a node decreases over time.
7. The calculation of similarities is usually done by finding Euclidean/Manhattan distance between the two vectors.
8. This process is repeated till all the output nodes have updated weight vectors associated with them, effectively forming a map of nodes separated by different distances w.r.t. to the winning output node.

The algorithm starts to converge after a certain number of iterations, after which the final values in the output nodes project the groups or patterns observed in the input data.

3.7 Evaluation Strategies

Before discussing evaluation strategies, it is important to note that this study has proposed a multistage process for the fast and effective analysis of prostate cancer images - a MapReduce based algorithm to scale the feature extraction step with dataset size and a SOM model for clustering the dataset. That being the case, it is only fitting to break down the evaluation in two parts.

The efficiency of the MapReduce approach to feature extraction can be easily evaluated by assessing the execution time under different scenarios. As mentioned previously, the MapReduce Architecture does tend to not show any improvements in performance when the input dataset is small enough to be held in memory, which seems to be the case with our 111 image dataset. It is likely that any other traditional approach for feature extraction will show faster execution than the MapReduce model. For a more fair evaluation, the dataset size must be increased and compared with currently available methods. There is a very easy solution to this problem - replicating the image dataset n times and executing different feature extraction algorithms on this dataset. This is just to compare the newly proposed model with the traditional ones and not to obtain the final feature vector - the final one will only contain 111 vectors for the 111 images.

Comparison with traditional approaches involves identifying alternate approaches to this process first. A rudimentary method for feature extraction would involve the use of an iterative program that loops over each image, reads it, performs some calculations and writes the extracted set of features into a file. Another way of doing this is the use of CITU (Computerized Image and Text Understanding) system, which is a computer program developed at Trinity College Dublin under the supervision of Professor Khurshid Ahmad, the details of which are described in the next chapter. This software has been used in previous projects as well and has shown promising results [29].

The performance scaling of the MapReduce based feature extraction system can also be analysed by observing execution times at a range of dataset sizes. This can be done by expanding the dataset on a binary scale, i.e., we can replicate the images on a scale of $(2^n) * 111$. This will provide us with daatsets of size 111,222,444,888, and so on. The slope of this graph can give sufficient information about how the system seems to scale with dataset size.

Coming to the clustering results obtained from the Self Organizing Maps, the best way to evaluate the results would be the formation of different clusters. Analyzing the different cluster patterns and checking the distance between clusters would give a general idea of the algorithm's understanding of the difference between healthy and unhealthy cells. Assessing how each of the features in the feature set affect the clustering performance would also give enough insight into the tuning of the model, as well as eliminating redundant features. The effects of the additional pre-processing done earlier can also be observed, any improvement or deterioration in cluster formation can be examined.

3.8 Conclusion

The above chapter has elaborated on the various steps involved in the pipeline of the proposed prostate tissue analysis model. The first two sections extensively discussed the background of the problem domain(3.2) and the data source and acquisition process. The steps taken to enhance and refine the image features are discussed next(3.3), followed by a detailed discussion on the type of features to be extracted(3.4). Next, the methods for feature extraction used in this study are explained, with a brief overview of the implementation design within our proposed model (section 3.5). Finally, the techniques to be used in this study for the clustering and analysis of the dataset (section 3.7) along with the evaluation methods(section 3.8) are elaborated on.

Chapter 4

Implementation and Evaluation

4.1 Introduction

In the previous chapter, several concepts, algorithms and frameworks pertaining to the approach used for prostate tissue analysis in this work were discussed. As such, this chapter aims to explain how each of these segments of the methodology fit into this work's proposed model, and how they work together to accomplish the research objectives outlined previously. The dataset used for the implementation of this Image Analysis System is first described. Then, the System Architecture of said model is discussed, followed by the tools and technology needed to implement the necessary steps. How these tools are leveraged for the purposes of this study is also discussed, be it coding paradigms or tool usage. This is followed by the evaluation of the results obtained, under different scenarios discussed under different case studies. The performance metrics are also included in the evaluation of the system. Lastly, the security and privacy concerns related to this system and data used are discussed, followed by a brief conclusion.

4.2 Dataset Description

As already discussed in section 3.2, the dataset used for this study consists entirely of high resolution Whole Slide Images of prostate tissue samples. To reiterate, these images were provided by Dr Aamir Ahmad, Head of Prostate Cancer Group at King's College London. There are a 111 images in total, each with dimensions 2000 x 2000 pixels and consuming 1.2 Megabytes of memory space. The data is completely unlabelled which means there is no known descriptors for any of the images, except for the information contained within it's pixels. Figure 4.1 shows a single image from the prostate tissue images dataset as an example for visual analysis.

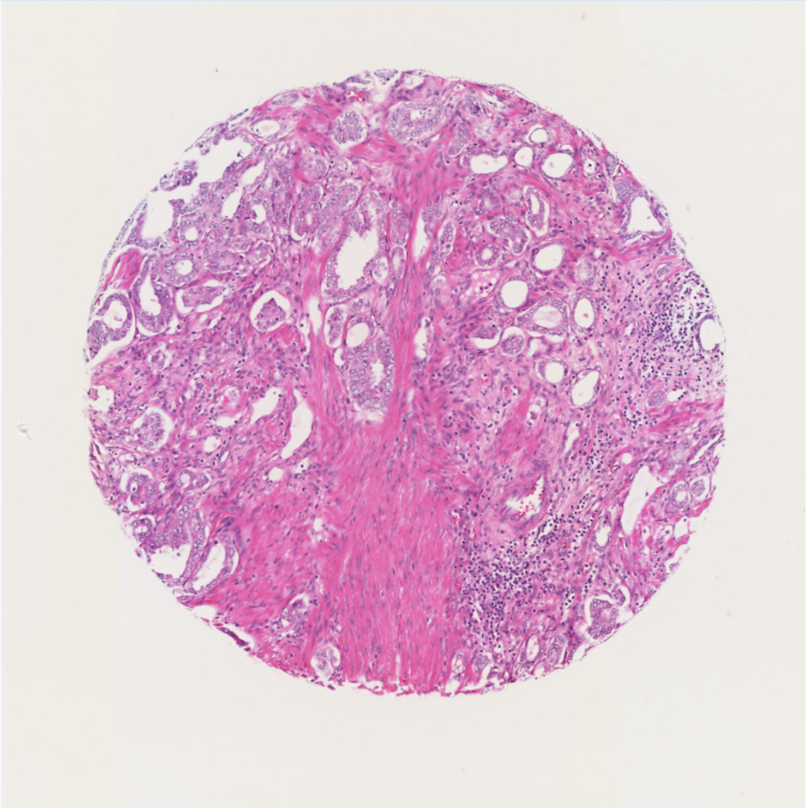


Figure 4.1: A sample image from the prostate cancer image dataset provided by King's College London

4.3 System Design and Technical Implementation

The section describes the design of the proposed system, detailing every tool used to implement this step by step. The details of the technical stack used, programming language and user interface description is also included.

4.3.1 System Design

This system's overall architecture is made up of a number of independent parts that communicate with one another to process the data as needed at that stage. Figure 4.2 illustrates a comprehensive system architecture diagram. The prostate tissue image dataset serves as the foundation for the overall processing, as can be seen from the figure.

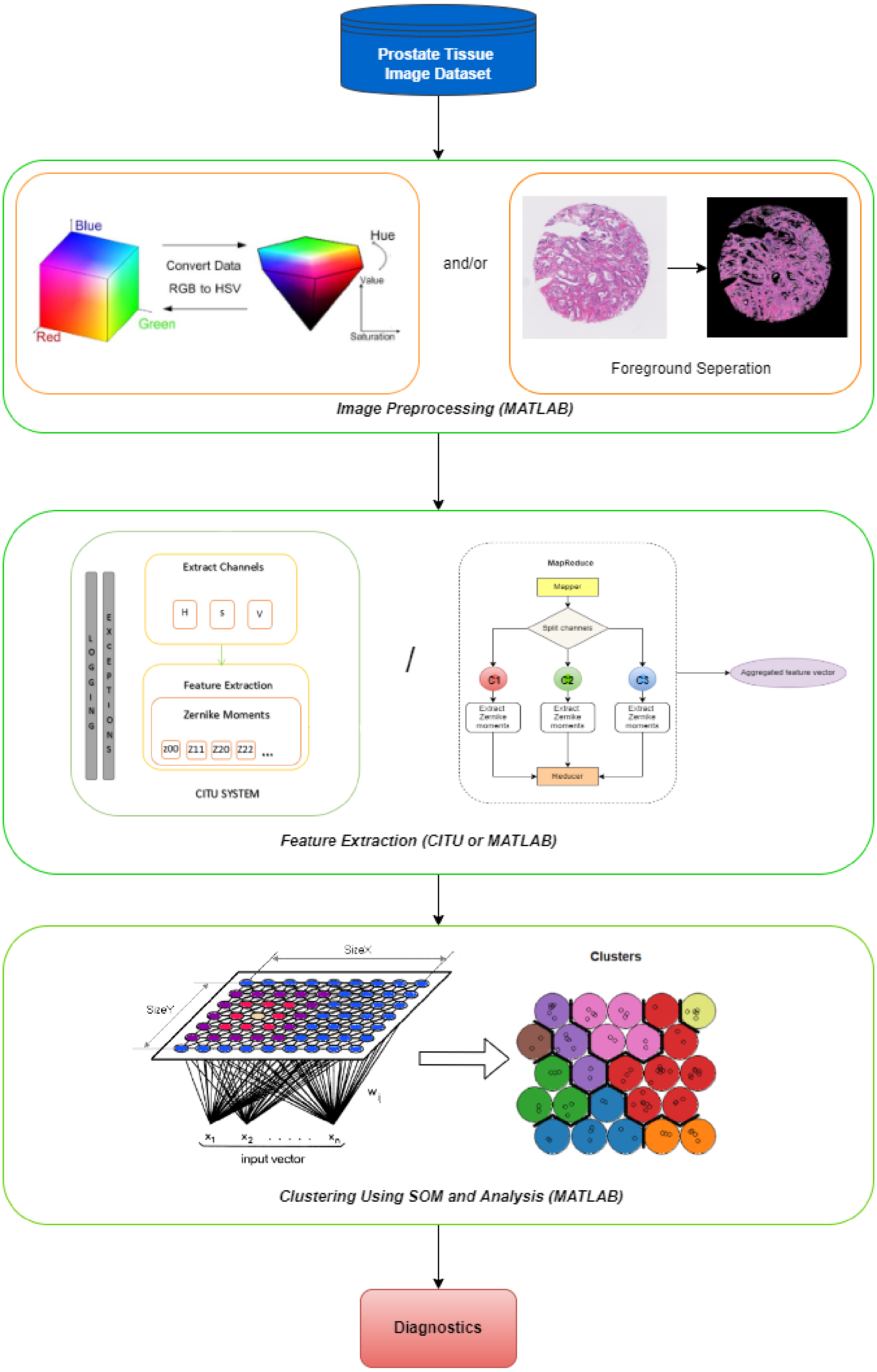


Figure 4.2: Architecture Diagram of Proposed System with tools and software used

The tools and softwares that aid in the implementation of each of these parts are explained in detail further in section 4.2.2.

4.3.2 Technical Stack

The different programs, tools, and technologies mentioned in the system design architecture are described in depth here. This section will examine the functionality, constraints, and design of the technologies employed in this work.

Image Pre-Processing Script and Tools

For the purposes of pre-processing, two separate scripts were written to translate the RGB channel encoding to HSV channel encoding and to separate foreground from the background of the image. Both of these operations were done on the MATLAB platform, and to apply these transformations on the entire dataset, an app called Image Batch Processor built into the MATLAB image processing toolkit was used.

The separation of foreground from the background in the images is slightly complex. This is done using color based segmentation, with the implementation of k-means clustering in MATLAB. First, the color space of the image is converted to the L*a*b color space from RGB. Doing this shows improved results in segmentation due to the fact that the L*a*b* color space separates image luminosity and color, making it easier to segment regions by color, independent of brightness of the pixels. The `rgb2lab()` method is used for this step. Then the image is segmented with the help of `imgsegkmeans()` method, where the number of clusters is specified as 2. Two clusters are specified because the foreground and background consist of two distinct colors and this should be sufficient to separate the two regions. The results of this process leaves us with two different image masks, one containing the foreground of the image and he other containing the background. The foreground mask is returned as the output image and the rest is discarded.

The tranformation from RGB to HSV is fairly simple, and is done by a function called `rgb2hsv()`, available in MATLAB that directly maps the red, green and blue color channels in an image to corresponding hue, saturation and value encoding.

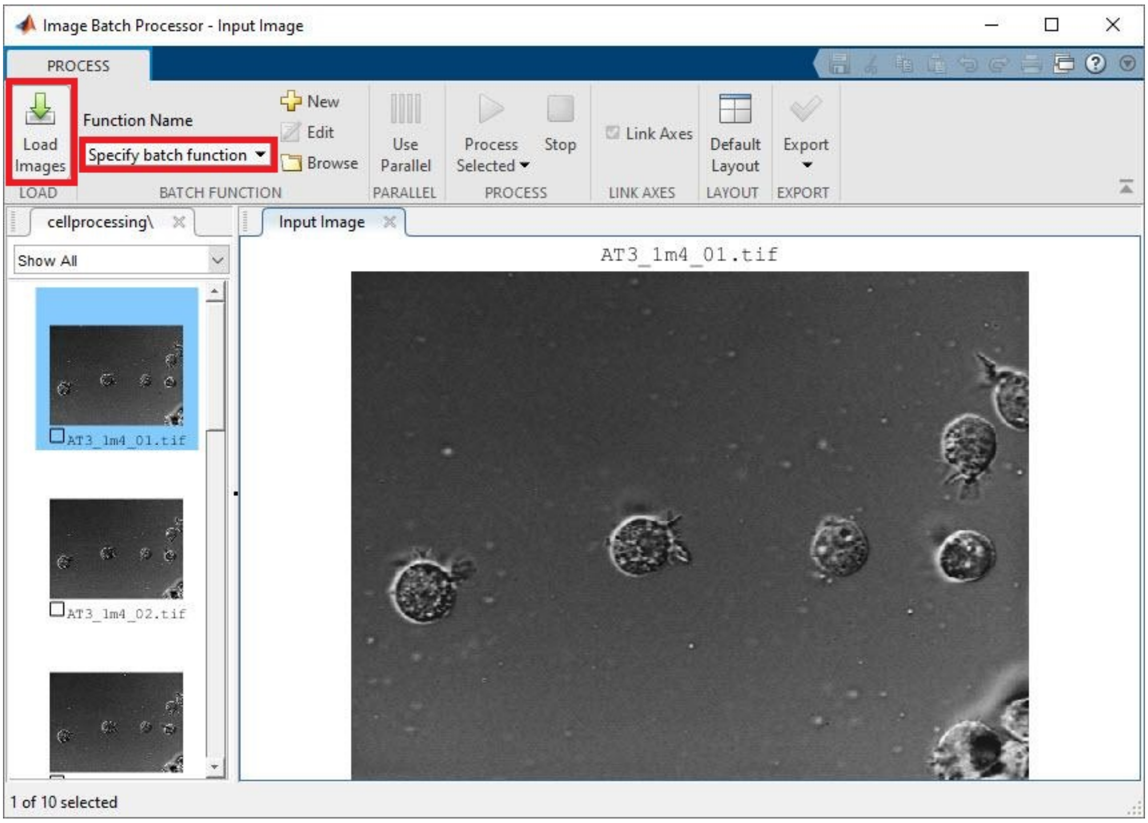


Figure 4.3: The Image Batch Processor in MATLAB. The image can be loaded directly from files or from the MATLAB workspace, and a specified function can be applied on an entire set of images at once.

Once the methods for the above two processes are written and evaluated on a single image sample, the entire dataset can be loaded into the Image Batch processor App, and desired transformation be applied on the images by specifying with function to be used as the batch function. These overall GUI of this app and the specified functionalities are highlighted in Figure 4.3.

CITU

CITU, which stands for Computerized Image and Text Understanding, is a Graphical User Interface application built on the Windows platform. It was developed under the supervision of Professor Khurshid Ahmad in Trinity College Dublin, for the purposes of image and text analysis, and to perform cross-modal analysis.

CITU was written on C++, and is packed with various functionalities for the analysis of images like segmentation using Otsu thresholding/ Watershed Algorithm, noise filtering and self-organising maps. For the examination of text based data the application supports features like TF-IDF, frequency computation and compound words. The tool also includes a cross-modal system to learn and form associations between image and textual features with the aid of annotated images.

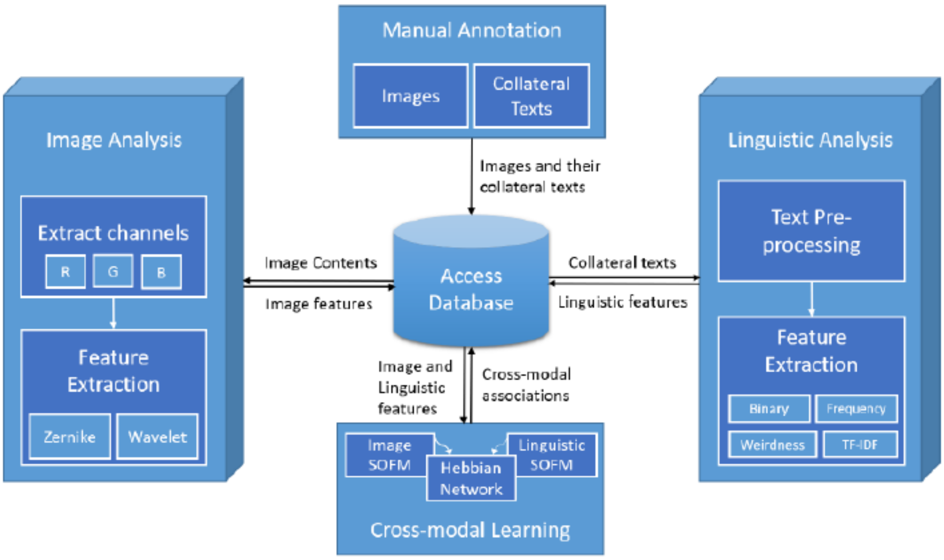
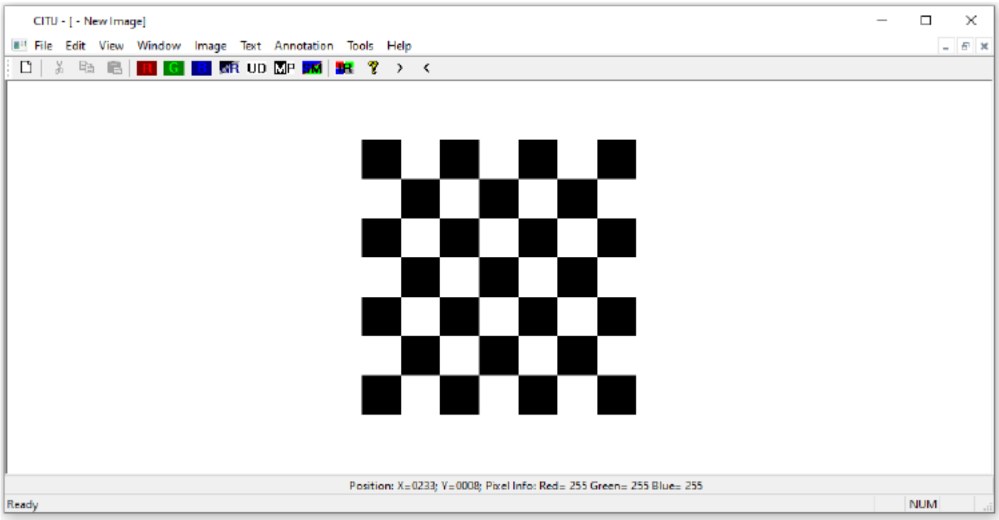


Figure 4.4: The cross-modal architecture of CITU facilitating image and text analysis

Amongst all these options, the one in focus in this study is feature extraction for image analysis. With this option CITU is able to extract Zernike Moments from any given image, after changing the RGB color encoding to the encoding of choice. In this case, the target encoding type can be chosen as HSV, so there is no need to separately apply the pre-processing script for HSV transformation in MATLAB. The CITU system allows the user to select an entire folder for feature extraction in one go, and the output of the entire set of files contained within the folder can be stored in a single output file specified in the program.



Position: X=0233; Y=0008; Pixel Info: Red= 255 Green= 255 Blue= 255

Figure 4.5: GUI of CITU application

MATLAB mapreduce

The MapReduce framework for extracting features proposed earlier has been implemented in the MATLAB programming environment for this study. MATLAB has provided the implementation of MapReduce in the form of the `mapreduce` function, which varies slightly from the traditional approach.

The MATLAB implementation of `mapreduce` requires that a `datastore` object for storing the input data be initialized first. This `datastore` object allows small chunks of the data to be stored in-memory at a time, after which it is passed on to the map and reduce phase. The map phase is encoded by the `map` function while the reduce phase is encoded by the `reduce` function. Both combined form the primary inputs to the `mapreduce` function.

Figure 4.6 shows the overall workflow of the `mapreduce` architecture and the different phases the data goes through after finally being translated into a meaningful output.

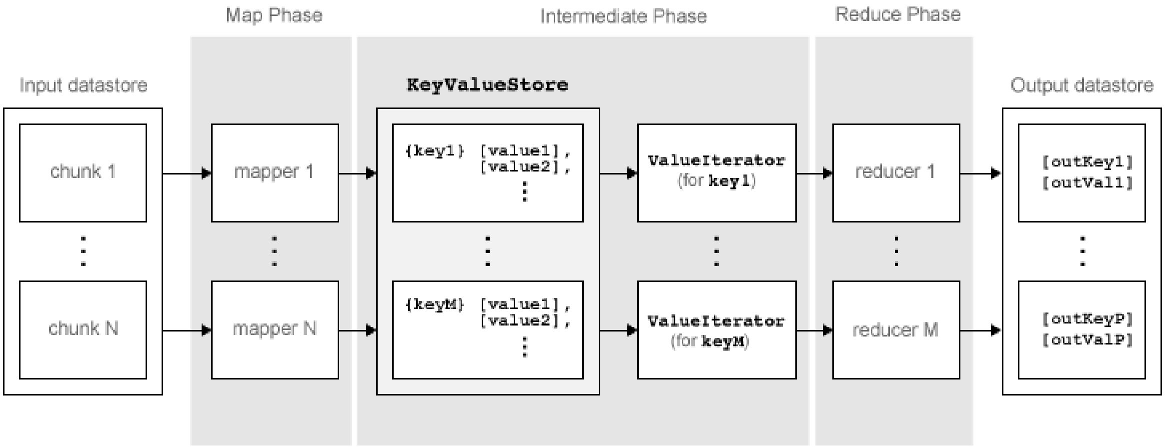


Figure 4.6: MATLAB mapreduce workflow diagram

The algorithm behind `mapreduce` follows the below steps to achieve parallelization of tasks while dealing with large datasets:

1. A chunk of data is read from the input datastore using the command `[data,info] = read(ds)`, where `ds` is the datastore object. This data is then passed on to the `map` phase.

2. In the map phase, the `map` function receives the read chunk of data and performs necessary computations to extract relevant information. It is then stored in an intermediate storage object called the `KeyValueStore` in the form of key value pairs. The number of chunks in the input datastore are indicative of the number of calls to the `map` function, and the operations are equally split across all instances of the mapper.

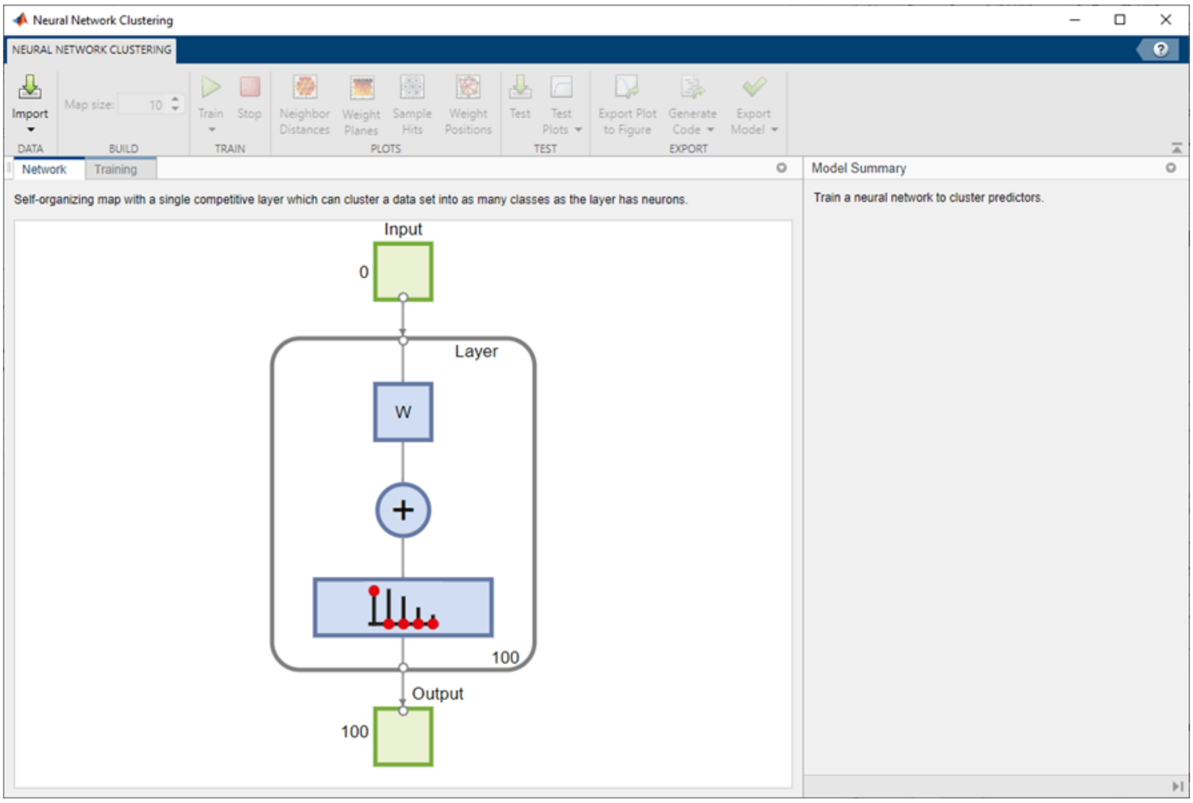
3. In the intermediate phase, the `mapreduce` algorithm groups together all the values in the `KeyValueStore` by unique keys.

4. These intermediate key value pairs are then passed on to the reduce phase, and calls are made to the `reduce` function for each unique key added by the `map` function. For each unique key, a `ValueIterator` object is passed to the reducer, which is used to iterate over all the values associated with that specific key.

5. The `reduce` function, after aggregating results from all these values, adds them to the output datastore using separate key value pairs.

Clustering Tools - MATLAB SOM

MATLAB comes prepackaged with several toolboxes for different scientific purposes - one of them is the Image Processing Toolkit mentioned earlier, which provides the Image Batch Processor App for processing multiple images at one go. Similarly, another toolbox has been provided by the MATLAB environment called the Deep Learning Toolbox, under which the Neural Network Clustering App is available. This can be directly launched from the prompt using the command `nctool`, or can be launched from the toolstrip under the Apps tab on top of the MATLAB window. Figure 4.7 shows the user interface of the Neural Network Clustering App along with the architecture of the neural network formed.



Self-organizing map with a single competitive layer which can cluster a data set into as many classes as the layer has neurons.

Train a neural network to cluster predictors.

Figure 4.7: MATLAB Neural Net Clustering App Window

The Neural Network Clustering App use the Self Organising Maps (SOM) architecture to solve clustering problems. The user can specify the number of dimensions of the SOM formed and also provide the input data for the clustering process using the import option. There is option to provide training and test data separately.

After training is complete, the clustering results can be evaluated using several different options, listed out below:

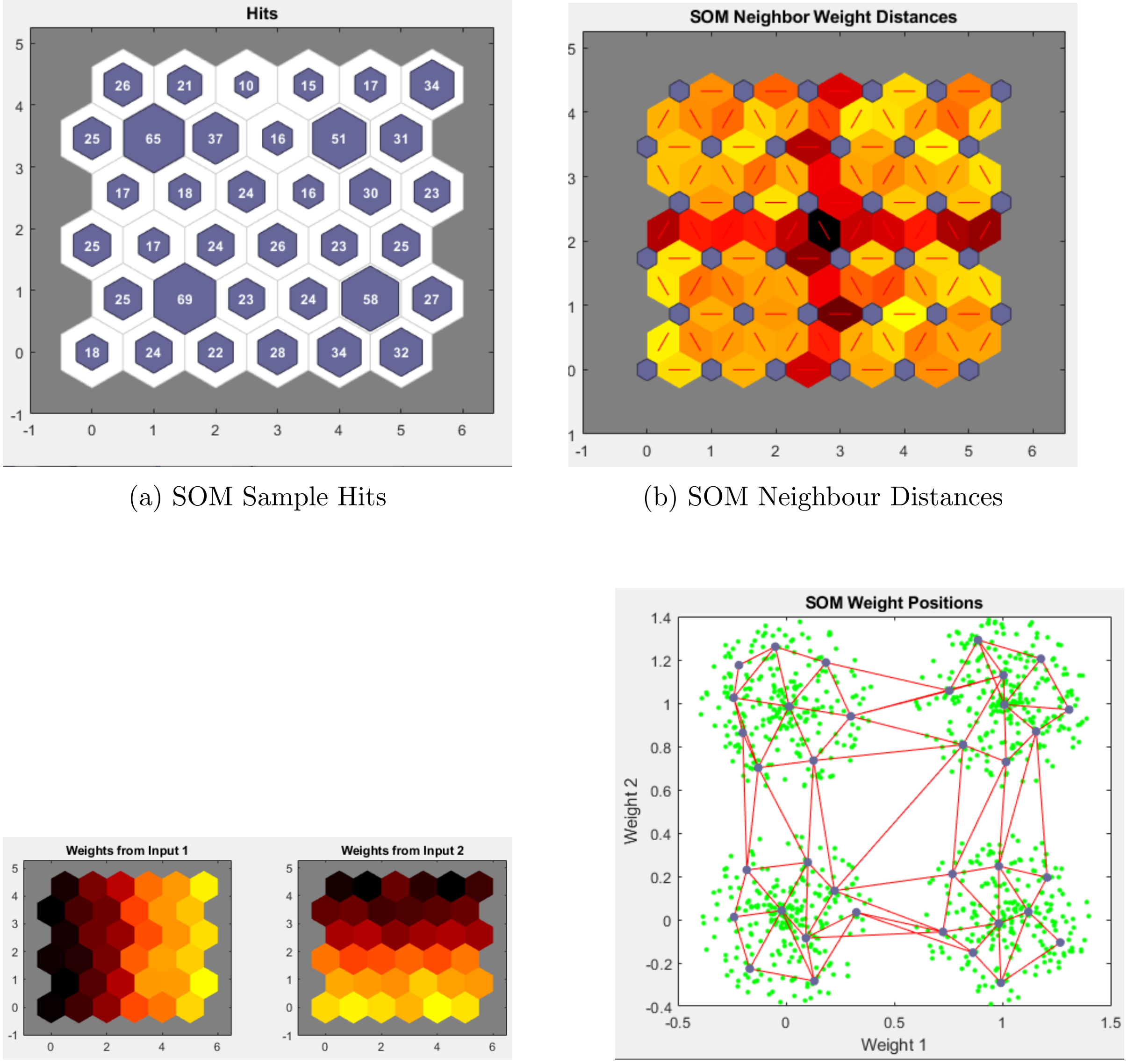
1. **SOM sample hits** - This plot is indicative of the cluster formations and the number of items that can be associated with each neuron is displayed in the form of hits. If a lot of items are grouped together it will be shown as a number of hits placed close together in the SOM output map.

2. **SOM neighbour distances** - The SOM neighbour weight distances show how close or far apart the neurons are with respect to each other after training. Each neuron is represented by a purple hexagon. Lighter colored connections are indicative of shorter distances and darker colored connections show that the neurons are far apart.

3. **SOM Input Planes** - The weights associated with each input can be evaluated separately by analysing this plot. The lighter colors indicate larger weights and darker ones indicate smaller weights.

4. **SOM Weight Positions** - This plot gives a visual insight into how the input vectors are spatially arranged with respect to weight vectors. High dimensional data is not suited to be examined with this method as there are multiple weight associations.

Figure 4.8 shows the different plots listed out above, taken from different studies as an example.



(a) SOM Sample Hits

(b) SOM Neighbour Distances

(c) SOM Weight Planes

(d) SOM Weight Positions

Figure 4.8: Examples of the plots available by MATLAB Neural Network Clustering App for the evaluation of clustering results. The images are examples obtained from other studies, none of them are indicative of the results in this study.

4.4 Case Studies and Evaluation

This section covers the performance evaluation of the proposed system under various scenarios. The performance of the MapReduce based feature extraction component and the SOM based clustering component are evaluated in different subsections for clarity. The limitations of the system are also explained in the last subsection.

4.4.1 Evaluation of Feature Extraction Methods

As explained in section 3.7, the most effective way to analyze the performance of the proposed MapReduce based feature extraction model is to check the execution time of the code. For a relative examination strategy, the performance times of traditionally used feature extraction methods are also included in this study. Traditional methods of feature extraction usually follow an iteration based approach which are often computationally expensive and time consuming. However, an important factor to be noted is that the performance of MapReduce based systems scales up with the size of the dataset, and performance improvement may not be significant when small amounts of data is being processed. To asses this aspect of the system, it is imperative that the evaluation be conducted with different dataset sizes with different approaches.

Keeping in line with the above considerations, two different iterative approaches are chosen for the comparative analysis of our proposed framework. One is a simple MATLAB based script, that iterates over the entire dataset of images, performs computations based on the mathematical concepts discussed in section 3.4.3 for the calculation of Zernike moments and returns the feature vector of each image to be written to a file. The other traditional approach to feature extraction us through the use of CITU - as it has been used in other studies previously and has shown good results. As for the analysis of the performance scaling claims of MapReduce, this can be easily evaluated by replicating the given dataset 10 times, giving us 1110 images and then using them as an input to the feature extraction model as explained in section 3.7. For a bit more perspective on computational performance, all these processes were run on a Windows based system with a 64 bit operating system, with 16GB RAM and an Intel Core I7, 4.20 GHz processor.

Table 4.1: Execution time of different approaches of feature extraction used in this study

Case	Approach	Platform	No. Images	Execution time (seconds)
1	Iterative	MATLAB	111	405.076235
2			1110	12054.8026
3		CITU	111	∼6000
4			1110	∼43200
5	MapReduce	MATLAB	111	616.894115
6			1110	5892.687563

Table 4.1 shows the cumulative results of all the different usecases possible with the different approaches, platforms and dataset sizes used in these experiments. As evident from these recordings, when processing a smaller dataset, i.e. 111 images, the basic iterative approach implemented in MATLAB performs the best, finishing execution in about 405 seconds or 6.75 minutes. The MapReduce based system does not perform better in this scenario, it in fact takes almost 200 seconds more than that needed in the iterative approach. The performance of CITU is not great in comparison to either of the other methods, it takes almost 2 hours to complete the extraction process for 111 images. However, when processing 1110 images, a considerable improvement in performance is noticed with the MapReduce model, it takes almost 6,162 seconds lesser than the MATLAB iterative approach to process the same dataset. CITU fares the worst this time as well taking almost 12 hours for the same task.

To visualise how the MapReduce framework's performance scales with an increase in dataset size, we can evaluate it's performance at different dataset sizes. This is best done with binary scaling, as this makes visual evaluation of performance easier by generating logarithmic plots of dataset size vs execution time. Binary scaling of data can be achieved by scaling the data up in terms of $(2^n) * 111$, where $n = 1, 2, 3, 4...$,etc. This will yield datasets of size 111, 222, 444 and so on. In this study, we have evaluated the execution performance of the proposed algorithm with datatset sizes 111, 222, 444, 888, 1776 and 3552.

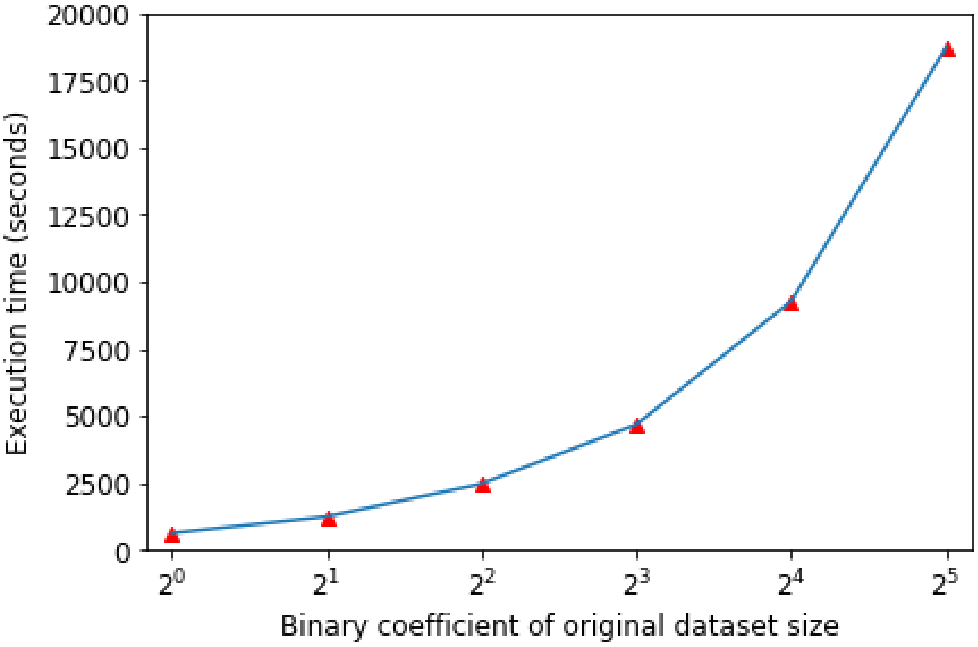


Figure 4.9: A plot depicting how MapReduce based system's performance scales up with a binary increase in dataset size

Figure 4.9 shows a plot of dataset size vs execution times on a logarithmic scale. It is evident from the figure that there is nearly a constant increase in the slope of this graph, which indicates that there is a linear increase in execution time with a binary increase in dataset size. This is further corroborated by table 4.2, which displays the points plotted in figure 4.9, along with the linear scale slope. The slope in these observations is seen to be almost constant.

Table 4.2: Execution times of MapReduce based feature extraction system with different dataset sizes, slope is the execution time divided by dataset size

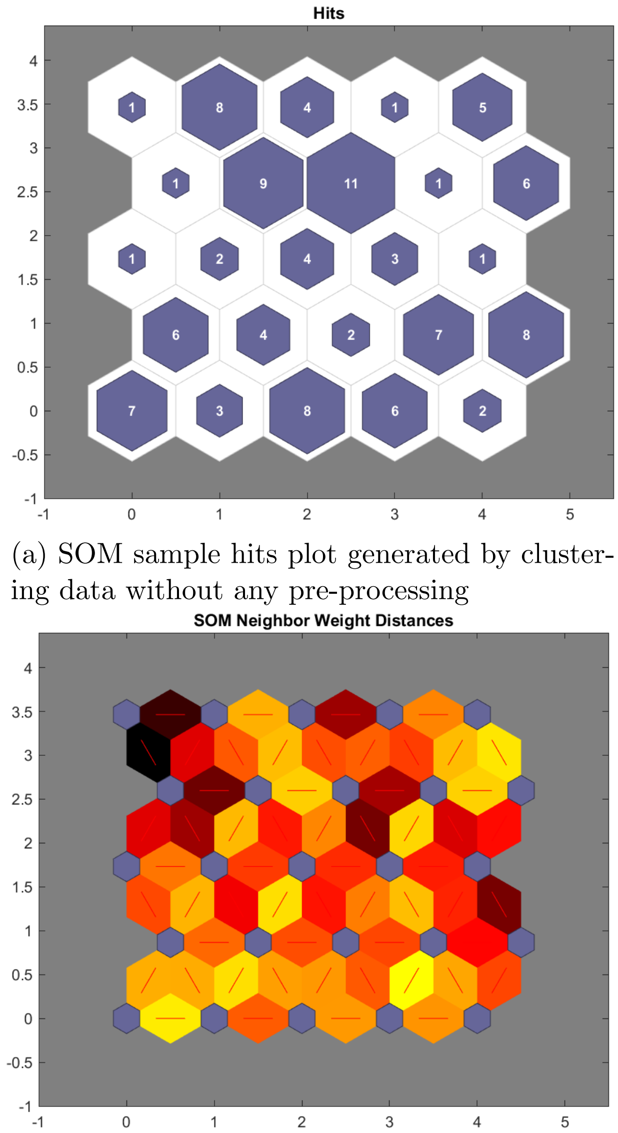
Dataset Size	Execution time (s)	Slope
111	616.894115	5.557605
222	1238.636091	5.579442
444	2462.822105	5.546897
888	4674.685357	5.264285
1776	9265.782278	5.21722
3552	18728.62827	5.272699

In Figure 4.9, it is evident that there is a progressive increase in the slope of the line plot. This is indicative of the fact that every time the dataset is doubled in size from what it was previously, the system extracts features much faster than before.

4.4.2 Clustering Evaluation

The evaluation of the clustering done on the dataset with the help of the Neural Network Clustering can be done by the various plots generated with this app, as explained in section 4.3. Initially, the Zernike Moments of the original dataset was taken, without applying any image pre-processing methods. These images included a lot of redundant information because of the presence of background pixels and noise. When the features taken from these images was passed to the clustering model as an input, it was observed that there was room from improvement - this will be explained later in this section. To further improve the clustering performance, image pre-processing was applied to the dataset, eliminating background pixels and enhancing the image by converting it to HSV encoding. The SOM map dimensions chosen for both these scenarios was chosen to be 5x5, this is because a map of smaller dimensions would lack the level of refinement needed to evaluate cluster formations and a larger dimension was not suitable for a dataset of this small a size, doing so would lead to the formation of very sparse clusters with no definite structure of pattern.

Figure 4.10a and Figure 4.10b are indicative of the clustering results of prostate tissue images without any pre-processing applied on them. They show the SOM sample hits plot and the SOM neighbour distances plot for these images. As it is evident in from the sample hits plot, there are very sparse clusters formed and different samples have been associated with several different clusters. There is no definitive distinction between the two target classes, i.e. sick and healthy prostate tissues. This is further corroborated by the neighbour distances plot, where the distances between neurons (purple hexagons) show no definitive patterns or ridges indicating an explicit boundary. To reiterate the reasoning behind this analysis, the distances between neurons are indicated by the color shared, darker colors indicate larger distances and lighter colors indicate shorter distances.

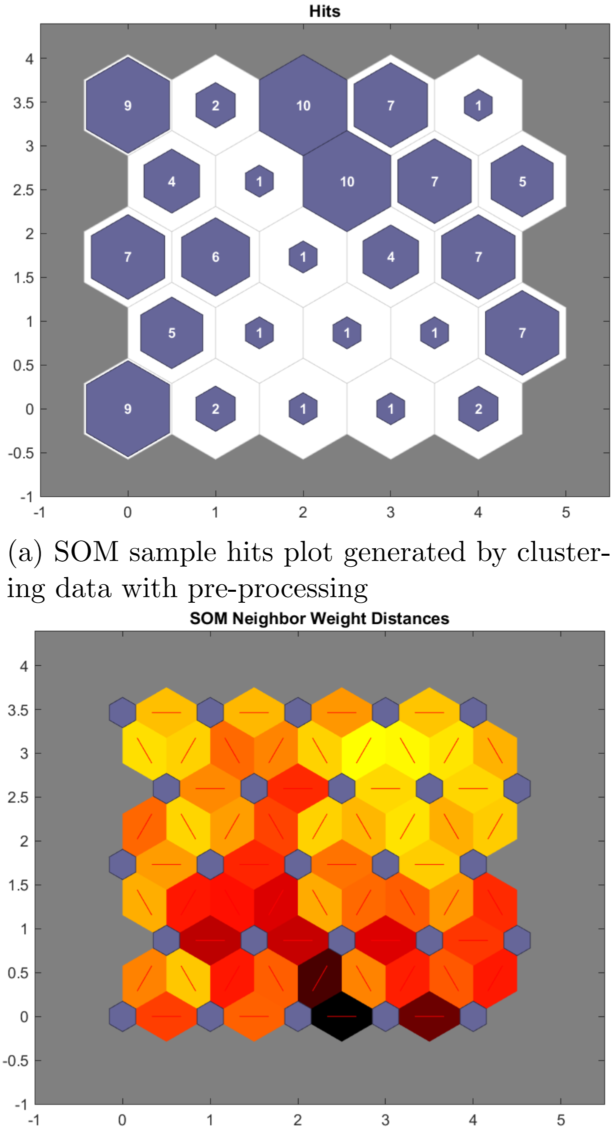


(a) SOM sample hits plot generated by clustering data without any pre-processing

(b) SOM neighbour distances plot generated by clustering data without any pre-processing

Figure 4.10: SOM evaluation plots generated by Neural Networks Clustering App for data without pre-processing

The results obtained after pre-processing the images are significantly different than the ones described before. Figure 4.11a shows the formation of two primary clusters, with small clusters with a very less number of hits indicating a boundary between the two clusters. To support this analysis, the neighbour distances plot (Figure 4.11b) is analysed, wherein the same boundary is observed between two clusters of closely spaced neurons. The clustering can be observed at the upper right-hand side of the map as well as on the left side of the map. This is separated by a darker colored region of neuron connections, indicating a divide between the two clusters.



(a) SOM sample hits plot generated by clustering data with pre-processing

(b) SOM neighbour distances plot generated by clustering data with pre-processing

Figure 4.11: SOM evaluation plots generated by Neural Networks Clustering App for data without pre-processing

The SOM weight positions plot was not taken into consideration for evaluation of this experiment because as mentioned earlier, the data is high dimensional containing 36 features (12 Zernike Moments for each color channel) and it is not possible to include all these weights in one plot. However, some insight can be gathered by looking at the SOM input weights plots, which shows how each feature contributes to weights associated with the SOM network.

Figure 4.12 and Figure 4.13 show the contribution of each input feature to the final SOM weight vector, in the case of images without pre-processing and in the case of images with pre-processing respectively. In both scenarios, it is observed that there are many inputs that contribute similar weights - like weights from inputs 13, 14, 15 and 17 in Figure 4.12 and weight from inputs 1, 3 and 10 in Figure 4.13. This is indicative of high correlation between the specified inputs in each case.

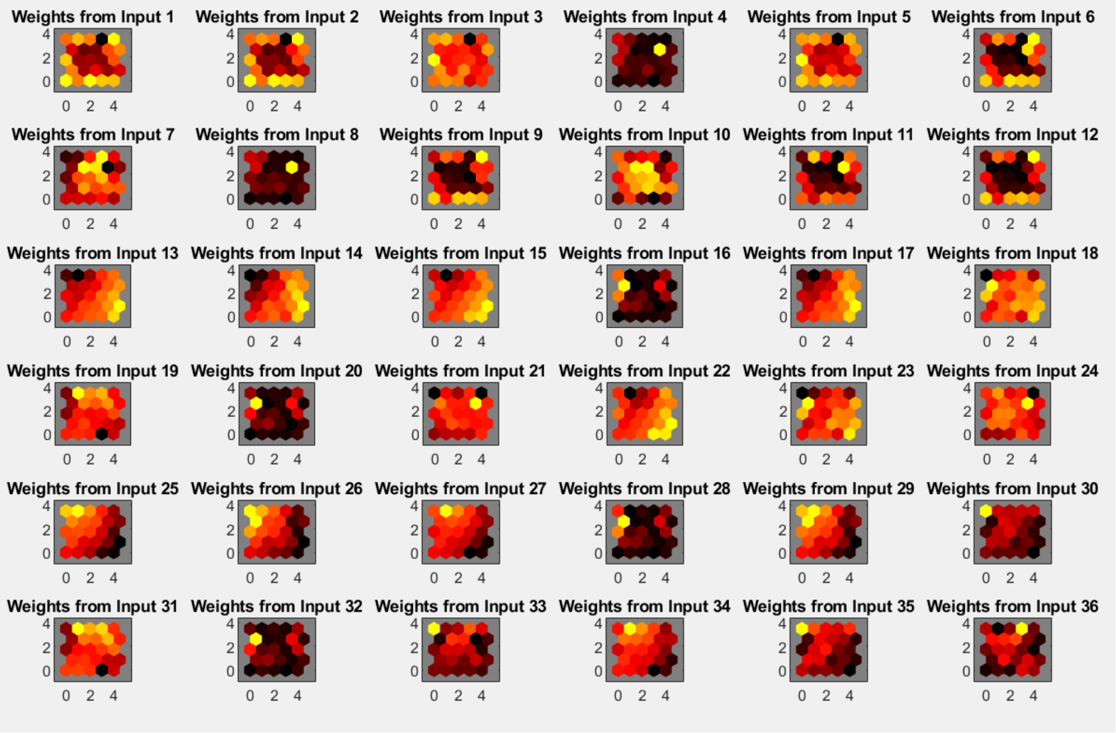


Figure 4.12: SOM sample hits plot generated by clustering data without any pre-processing

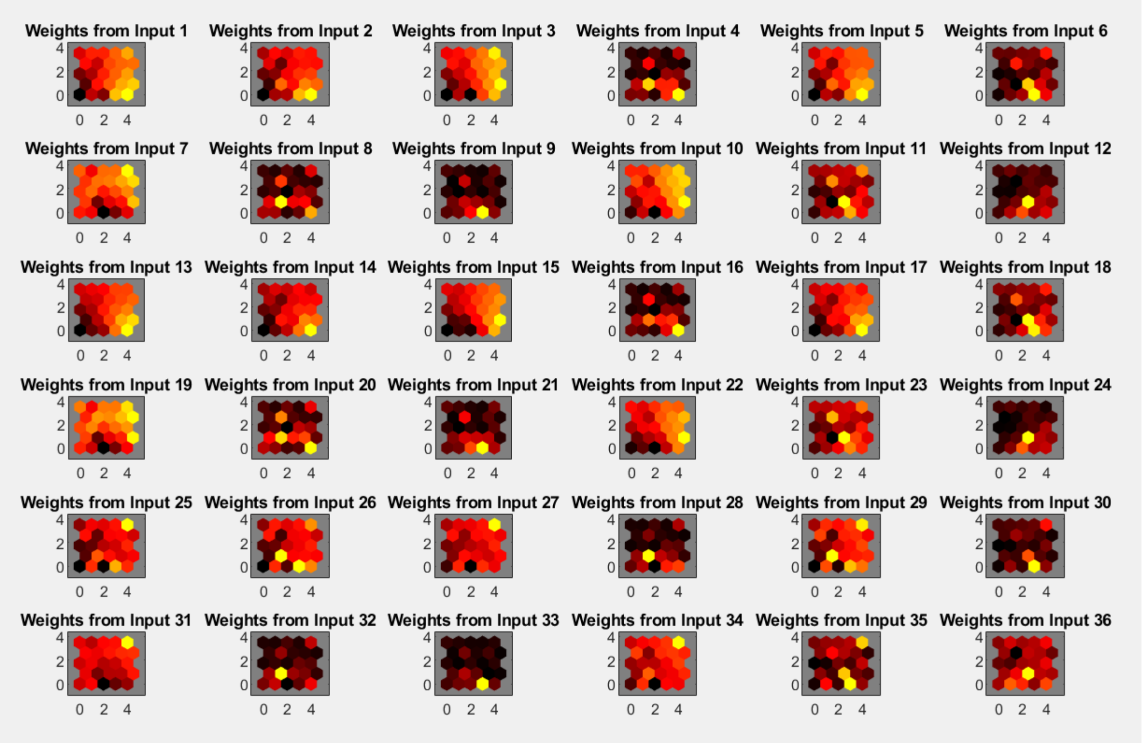


Figure 4.13: SOM neighbour distances plot generated by clustering data with pre-processing

4.5 Security and Privacy Concerns

As with any system working with several libraries, third party softwares and tools, this system is also susceptible to security threats and vulnerabilities. The data used in this study also comes under scrutiny as it involves medical imagery obtained from real individuals. This section addresses all these concerns and highlights any potential issues.

4.5.1 Security Concerns

This study has come up with a MapReduce based system for the optimized extraction of quantifiable features from high resolution images. While this has several benefits in terms of performance, there are certain issues that need to be taken care of. In MapReduce, there is no authentication between master and slave daemons[45]. As the computations are distributed overe several clusters, each cluster is vulnerable to attacks and data breaches. It is also susceptible to Denial of Service(DoS) attacks and SQL injection by malicious requests [45]. This can be a serious threat as dealing with medical and patient data is supposed to be extremely secure to protect personal data. The can be dealt with by using a Secure MapReduce architecture, as proposed in [46] and [47].

This study also uses several third party softwares and libraries, most of which come prepackaged with MATLAB. Since MATLAB has been around for quite sometime and has an active user feedback forum, most of the vulnerabilities are identified and addressed relatively fast. However, this study also uses CITU system, which is relatively new and has not undergone many code revisions as it was developed for the specific research purposes of an organisation. It uses libraries like OpenCV which has had several issues with buffer overflows in the past (CVE-2019-5063 and CVE-2019-5064). Care must be taken to update dependencies over time with tools like this.

Recent studies have brought into light a different type of threat that image classification systems face - this is done by the use of adversarial patches [48]. Simply put, adversarial patches are minute perturbations made to the image's pixels used for training or testing, and they can impair an image recognition/classification system's capacity to learn. They are an all-encompassing and reliable method for taking on any classification software that uses images as an input. Sufficient caution should be exercised to ensure that the data fed into these systems is not corrupted by patches of this kind.

4.5.2 Privacy and Ethical Concerns

As this study deals with the analysis of tissue imagery obtained by scanning samples taken from individuals, there may be several concerns raised about the privacy of the data collected. As mentioned earlier, the image dataset used in this work was provided by Dr. Aamir Ahmad, Head of the Prostate Cancer Group at King's College London. The tissue samples were collected at the Department of Pathology, Portuguese Oncology Institute, Porto, Portugal, Department of Pathology and Molecular Immunology, Abel Salazar Institute of Biomedical Sciences, University of Porto, Porto, Portugal, by Professor Rui Henrique. Their use was approved for research purposes by the Institutional Review Board. All the samples used in this study are anonymized, i.e., no personal information related to any of the individuals from whom the tissue was collected has been included in the dataset.

There are no ethical concerns as such since no actual human tissue was used, only their images were utilized for the purposes of this study.

4.5.3 Conclusion

This chapter has extensively covered the implementation and evaluation details of the proposed system. Section 4.2 included a brief recap about the input dataset and it's sources, section 4.3 covered the details of the technical stack utilized to implement and test this systems, section 4.4 covered the results and evaluation of both the feature extraction techniques and clustering techniques proposed in this study and finally the security and privacy concerns pertaining to this study were discussed in section 4.5.

Chapter 5

Conclusion and Future Work

This chapter puts forth the concluding points of the work undertaken in this dissertation. Additionally, the limitations of the research and experiments conducted in this study are highlighted, with suggestions for improvement in future iterations of this work.

5.1 Conclusion

The field of medical image analysis is at its nascent stage, needing improvements on several fronts. This work tries to address the various issues in the current pipeline of image analysis for pathology by proposing a MapReduce and Clustering based approach to evaluate the images for diagnostic purposes. The image is first turned into a numerical feature vector using the mathematical transformations first described by physicist Fritz Zernike. This leaves us with global descriptors of the image, calculated separately over the three colour channels, called Zernike Moments. The traditional methods for performing the calculations needed to extract these features have been deemed inefficient and time-consuming, so a parallel computing approach based on MapReduce has been proposed for these purposes. The Zernike Moments extracted from the images gives us features that are unaffected by translational, rotational or position changes. They also happen to be orthogonal, eliminating any redundancy or chances of bias. Once these features are extracted, they are passed as an input to an unsupervised clustering model based on the Kohonen Maps, also known as Self Organising Maps. The unsupervised method of clustering is the preferred way of analysis due to the lack of labels in the used dataset. Self Organising Maps operate by creating low-dimensional discretized projections of the input data into maps and learn to create associations between each of the inputs passed. Thus, using Zernike Moments of an image, the SOM network is able to cluster the different images into different clusters indicating that the system is able to distinguish between sick and healthy cells. The proposed MapReduce-based feature extraction method performs much better than traditional approaches like CITU and iterative programs, cutting down processing times by more than half in most scenarios. The performance of this system also scales linearly with a binary increase in data. As for the clustering methods used, with the right pre-processing of input data, the algorithm is able to distinguish between the two groups of tissues indicating different health conditions.

This study presents work that can potentially be used in the field of histopathology to evaluate and assess biological samples obtained from patients in a relatively fast and economical manner. This pushes the field of medicine a step further in digitising the infrastructure, which is much needed to provide speedy and affordable healthcare to all. However, the proposed system on its own cannot be used for the diagnosis of life-threatening and chronic diseases like cancer - the accuracy levels of Machine Learning systems are not yet on par with that of human experts. Caution should be exercised when analysing samples with such systems, and they should be used for screening purposes at best, always supported by the expert opinions provided by professionals.

5.2 Future work

Post assessing the performance of the proposed architecture via various experiments and evaluation methods, it is observed that there are certain limitations that the proposed approach of medical analysis faces. There is certainly room for improvement in this model, be it in the data storage aspect or classification aspect. Some of the potential improvements to the system are listed down below based on the shortcomings of the system.

1. The proposed model of image analysis can certainly benefit from the use of a much larger dataset to train the Clustering Model. The SOM network in this study was not able to present a clear division between the two final clusters meaning there was some overlap in the samples. A larger number of samples will lead to better adjustments of weight in the network and show better clustering.

2. If used for pathological purposes, this system will most likely encounter unlabelled data, which was used in this study as well. However, it is better to have labelled data for training purposes so that the accuracy of the system can be evaluated before using it for diagnosis.

3. This study proposes a MapReduce-based feature extraction model that has the ability to work with cloud-based storage systems. However, most of the experiments

performed were with datasets that could be held in memory. It would be interesting to see how this approach scales with large distributed storage datastores like Hadoop and Azure.

4. Multi-modal systems like CITU can be used to provide an additional set of features apart from the visual characteristics of the image for better classification results. This will help overcome the challenges posed by errors and noise introduced in images by imaging devices.

The above suggestions are purely based on the weaknesses identified in this study, and they can be rectified in future iterations of this work. However, this is not an exhaustive list, and future developments in science and technology may build upon these findings in a way unforeseen before.

Bibliography

[1] P. Rawla, “Epidemiology of prostate cancer,” *World journal of oncology*, vol. 10, no. 2, p. 63, 2019.

[2] I. C. S. Ireland, *Prostate cancer*, Aug. 2021. [Online]. Available: `https://www.cancer.ie/cancer-information-and-support/cancer-types/prostate-cancer`.

[3] M. Quinn and P. Babb, “Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. part i: International comparisons,” *BJU international*, vol. 90, no. 2, pp. 162–173, 2002.

[4] G. Draisma, R. Etzioni, A. Tsodikov, *et al.*, “Lead time and overdiagnosis in prostate-specific antigen screening: Importance of methods and context,” *Journal of the National Cancer Institute*, vol. 101, no. 6, pp. 374–383, 2009.

[5] R. Etzioni, D. F. Penson, J. M. Legler, *et al.*, “Overdiagnosis due to prostate-specific antigen screening: Lessons from us prostate cancer incidence trends,” *Journal of the National Cancer Institute*, vol. 94, no. 13, pp. 981–990, 2002.

[6] J. R. Prensner, M. A. Rubin, J. T. Wei, and A. M. Chinnaiyan, “Beyond psa: The next generation of prostate cancer biomarkers,” *Science translational medicine*, vol. 4, no. 127, 127rv3–127rv3, 2012.

[7] N. Mehta, A. Raja’S, and V. Chaudhary, “Content based sub-image retrieval system for high resolution pathology images using salient interest points,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 3719–3722.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[9] M. Heidari, S. Mirniaharikandehei, G. Danala, Y. Qiu, and B. Zheng, "A new case-based cad scheme using a hierarchical ssim feature extraction method to classify between malignant and benign cases," in *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, SPIE, vol. 11318, 2020, pp. 309–315.

[10] H. Vo, J. Kong, D. Teng, *et al.*, "Mareia: A cloud mapreduce based high performance whole slide image analysis framework," *Distributed and parallel databases*, vol. 37, no. 2, pp. 251–272, 2019.

[11] S. M. Mahmoud and R. S. Habeeb, "Analysis of large set of images using mapreduce framework," *International Journal of Modern Education and Computer Science*, vol. 11, no. 12, p. 47, 2019.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, Ieee, vol. 2, 1999, pp. 1150–1157.

[13] S. Benlakhdar, M. Rziza, and R. O. H. Thami, "A robust model using sift and gamma mixture model for texture images classification: Perspectives for medical applications," *Biomedical and Pharmacology Journal*, vol. 13, no. 4, pp. 1659–1669, 2020.

[14] Y. Kumar, A. Aggarwal, S. Tiwari, and K. Singh, "An efficient and robust approach for biomedical image retrieval using zernike moments," *Biomedical Signal Processing and Control*, vol. 39, pp. 459–473, 2018.

[15] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[16] N. Bastanfard Azam & Ahangar, "A comparison between sift descriptor and zernike moments feature extraction on geometric shapes," in *1st international Conference on New perspective in Electrical and computer Engineering*, 2010.

[17] K. Wu, C. Garnier, J.-L. Coatrieux, and H. Shu, "A preliminary study of moment-based texture analysis for medical images," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, 2010, pp. 5581–5584.

[18] Z. Iscan, Z. Dokur, and T. Ölmez, "Tumor detection by using zernike moments on segmented magnetic resonance brain images," *Expert systems with applications*, vol. 37, no. 3, pp. 2540–2549, 2010.

[19] C. Zheng, A. Long, Y. Volkov, A. Davies, D. Kelleher, and K. Ahmad, "A cross-modal system for cell migration image annotation and retrieval," in *2007 International Joint Conference on Neural Networks*, IEEE, 2007, pp. 1738–1743.

[20] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[21] S. Mishra and M. Panda, “Medical image retrieval using self-organising map on texture features,” *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 359–370, 2018.

[22] T. Araújo, G. Aresta, E. Castro, *et al.*, “Classification of breast cancer histology images using convolutional neural networks,” *PloS one*, vol. 12, no. 6, e0177544, 2017.

[23] N. Hegde, J. D. Hipp, Y. Liu, *et al.*, “Similar image search for histopathology: Smily,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[24] K. Shrivastava, N. Gupta, and N. Sharma, “Medical image segmentation using modified k means clustering,” *International Journal of Computer Applications*, vol. 103, no. 16, 2014.

[25] A. Rana, G. Yauney, A. Lowe, and P. Shah, “Computational histological staining and destaining of prostate core biopsy rgb images with generative adversarial neural networks,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2018, pp. 828–834.

[26] E. A. Krupinski, J. P. Johnson, S. Jaw, A. R. Graham, and R. S. Weinstein, “Compressing pathology whole-slide images using a human and model observer evaluation,” *Journal of pathology informatics*, vol. 3, no. 1, p. 17, 2012.

[27] M. M. Dundar, S. Badve, G. Bilgin, *et al.*, “Computerized classification of intraductal breast lesions using histopathological images,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 7, pp. 1977–1984, 2011.

[28] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, “Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives,” *Journal of pathology informatics*, vol. 9, no. 1, p. 40, 2018.

[29] A. Verma, “Classification of medical images using artificial neural network,” M.S. thesis, 2020.

[30] A. T. Feldman and D. Wolfe, “Tissue processing and hematoxylin and eosin staining,” in *Histopathology*, Springer, 2014, pp. 31–43.

[31] A. J. Symes, M. Eilertsen, M. Millar, *et al.*, “Quantitative analysis of btf3, hint1, ndrg1 and odc1 protein over-expression in human prostate cancer tissue,” *PloS one*, vol. 8, no. 12, e84295, 2013.

[32] K. Cao, C. Arthurs, A. Atta-Ul, *et al.*, "Quantitative analysis of seven new prostate cancer biomarkers and the potential future of the 'biomarker laboratory'," *Diagnostics*, vol. 8, no. 3, p. 49, 2018.

[33] C. Arthurs, B. N. Murtaza, C. Thomson, *et al.*, "Expression of ribosomal proteins in normal and cancerous human prostate tissue," *PLoS One*, vol. 12, no. 10, e0186047, 2017.

[34] M. Akbari, M. Mohrekesh, K. Najariani, N. Karimi, S. Samavi, and S. R. Soroushmehr, "Adaptive specular reflection detection and inpainting in colonoscopy video frames," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 3134–3138.

[35] D. D. Patil and S. G. Deore, "Medical image segmentation: A review," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 1, pp. 22–27, 2013.

[36] H. Ng, S. Ong, K. Foong, P.-S. Goh, and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *2006 IEEE southwest symposium on image analysis and interpretation*, IEEE, 2006, pp. 61–65.

[37] J. Katkar, T. Baraskar, and V. R. Mankar, "A novel approach for medical image segmentation using pca and k-means clustering," in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, IEEE, 2015, pp. 430–435.

[38] M. Zhao, Q. Chai, and S. Zhang, "A method of image feature extraction using wavelet transforms," in *International Conference on Intelligent Computing*, Springer, 2009, pp. 187–192.

[39] K. H. Ghazali, M. F. Mansor, M. M. Mustafa, and A. Hussain, "Feature extraction technique using discrete wavelet transform for image classification," in *2007 5th Student Conference on Research and Development*, IEEE, 2007, pp. 1–4.

[40] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, vol. 8, no. 2, pp. 179–187, 1962.

[41] P.-T. Yap, X. Jiang, and A. C. Kot, "Two-dimensional polar harmonic transforms for invariant image representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1259–1270, 2009.

[42] C. Singh *et al.*, "Local and global features based image retrieval system using orthogonal radial moments," *Optics and Lasers in Engineering*, vol. 50, no. 5, pp. 655–667, 2012.

[43] C. Singh and P. Sharma, “Performance analysis of various local and global shape descriptors for image retrieval,” *Multimedia systems*, vol. 19, no. 4, pp. 339–357, 2013.

[44] A. Khotanzad and Y. H. Hong, “Invariant image recognition by zernike moments,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[45] G. S. Bhathal and A. Singh, “Big data: Hadoop framework vulnerabilities, security issues and attacks,” *Array*, vol. 1, p. 100 002, 2019.

[46] P. Jain, M. Gyanchandani, and N. Khare, “Enhanced secured map reduce layer for big data privacy and security,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–17, 2019.

[47] E. Fabiano, M. Seo, X. Wu, and C. C. Douglas, “Opendbddas toolkit: Secure mapreduce and hadoop-like systems,” *Procedia Computer Science*, vol. 51, pp. 1675–1684, 2015.

[48] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017.

Appendix

...