

# **Classification of Medical Images using Artificial Neural Network**

**Akash Verma**

**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

September 2020

# 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.

---

Akash Verma

September 11, 2020

## Permission to Lend and/or Copy

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

---

Akash Verma

September 11, 2020

# Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor Khurshid Ahmad whose passion for exploring, patience while explaining concepts and vast knowledge of different fields of study have made this gratifying journey possible and enjoyable.

As importantly, I would like to thank Dr. Aamir Ahmed, Head of Stem Cell and Prostate Cancer Group, King's College London, and Rui Henrique, Professor at Department of Pathology and Molecular Immunology, University of Porto, for providing data and sharing extremely valuable insights throughout the research.

A special thanks to my mother Manjula V, my father Dhananjay V, and my sister Urvashi V for their endless love, support, and encouragement to pursue my dreams.

Finally, I would like to thank all my friends for being there to support and keeping me motivated. As well as Rosemarie Power, Systems Administrator, Computer Science, for providing technical support during remote work.

AKASH VERMA

*University of Dublin, Trinity College  
September 2020*

# Classification of Medical Images using Artificial Neural Network

Akash Verma, Master of Science in Computer Science  
University of Dublin, Trinity College, 2020

Supervisor: Professor Khurshid Ahmad

Medical imaging is a rapidly developing field and is complemented by the enhancement in image processing techniques including image enhancement, recognition, and analysis. The images taken provides insight into the inner structure of the body which can be used for medical as well as a scientific study, disorder identification, and treatment. Apart from its useful applications they have several other benefits like faster and economical processing, easier storage and transfer of data, and allowing enhanced manipulation of data. Advancements made in medical imaging over the last two decades have created unprecedented opportunities for new diagnostic approaches by utilizing the interdisciplinary field of medical image processing. In this work, various medical images of different types, acquired from different sources, are analyzed using an artificial neural network that utilizes a global image representation technique to denoise the confounded image data and provide a set of the image feature vector for analysis. The system makes use of clustering to provide a visualization of this similarly arranged data. The approach taken provides a generic solution that can be applied to an even wider medical image dataset. The practical application of this work is in the field of pathology where it can be used to provide a preliminary analysis of the dataset or give a second opinion to pathologists. The system was tested using 3 datasets of different sizes and dimensions achieving a minimum accuracy of 76%. This work showcases the potential that machine learning can have in the field of medical imaging and analysis.

# Summary

Image processing techniques are used to extract some useful information or to just enhance the quality of images being analyzed by performing various operations. These operations are a type of signal processing step where the input is an image and the extracted features or characteristics associated with the image is the calculated output. The complete process can basically be divided into three steps:

- Using Image acquisition to import image datasets.
- Manipulate and analyze the chosen image dataset.
- Present image analysis report or the altered image dataset as part of the output.

The use of image processing has been increasing exponentially over the last decade and has seen some tremendous growth. However, the growth in medical image processing, which is part of image processing, hasn't been that extreme due to the limitations associated with acquiring medical image data. Although where possible, the medical image processing techniques have made clinical diagnosis treatment and protocols more accurate and quite efficient.

In this work, a similar goal of finding an efficient way of classifying these medical images for diagnosis purpose is intended. To achieve this, the approach can broadly be divided into the steps mentioned above, where the process begins with image acquisition (in this case approx. 1650 images), detailing the information on images acquired and challenges associated with them. The second step of analysis is further broken

down into multiple steps of pre-processing, representation, and clustering, as these medical images are quite complex and require a well analyzed technique to gather the relevant information while ignoring any noise. A set of up to 12 input vectors are extracted and passed for analysis where images get clustered based on their similarity with neighboring nodes. Finally, this network of nodes is analyzed using test data to generate the analysis report that finds the accuracy of the system and visualize the output to know the distance between the various clusters.

The approach described is implemented using tools, such as Zen Lite and CITU, and technologies, like python and MATLAB. The overall architecture is described in a pipeline diagram that explains the flow of data and the evaluation method used. Furthermore, this work is conducted with inputs, suggestions, and subjective evaluation by subject matter experts from the medical domain to verify the produced results and provide comparison with the current state of the art approach for such medical images analysis.

# Contents

<b>Acknowledgments</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Summary</b>	<b>v</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Figures</b>	<b>xi</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Problem Definition . . . . .	5
1.2 Contributions . . . . .	7
1.3 Structure of the Dissertation . . . . .	8
<b>Chapter 2 Motivation and Literature Review</b>	<b>9</b>
2.1 Motivation . . . . .	9
2.2 Literature Review . . . . .	11
2.2.1 Imaging in Histopathology . . . . .	11
2.2.2 Image Feature Vectors . . . . .	12
2.2.3 Image representation using moments . . . . .	13
2.2.4 Learning technique in medical images . . . . .	13
2.3 Conclusion . . . . .	14
<b>Chapter 3 Method</b>	<b>16</b>
3.1 Introduction . . . . .	16



3.2	Knowledge and Data Acquisition . . . . .	18
3.2.1	Knowledge Acquisition . . . . .	18
3.2.2	Data Acquisition . . . . .	21
3.3	Preprocessing of Images . . . . .	23
3.4	Represent Image Using Moments . . . . .	26
3.4.1	Image Moments . . . . .	26
3.4.2	Zernike Moments . . . . .	27
3.5	Classification and Evaluation Methods . . . . .	32
3.5.1	Learning Approach for Medical Image Analysis . . . . .	35
3.5.2	Self Organizing Map . . . . .	36
3.5.3	Evaluation of Self Organizing Map . . . . .	39
3.6	Conclusion . . . . .	40
<b>Chapter 4 Implementation and Case Studies</b>		<b>41</b>
4.1	Introduction . . . . .	41
4.2	Dataset Description . . . . .	41
4.3	Technical Implementation and System Design . . . . .	43
4.3.1	System Architecture . . . . .	43
4.3.2	Technical Stack . . . . .	43
4.4	Case Studies and Discussion . . . . .	51
4.4.1	Evaluation of Endoscopy Images . . . . .	51
4.4.2	Evaluation of Lymph Node Images . . . . .	57
4.4.3	Evaluation of Prostate Tissue Images . . . . .	61
4.5	System Performance . . . . .	65
4.6	Security and Privacy Concerns . . . . .	65
4.6.1	Security Concerns . . . . .	65
4.6.2	Privacy Concerns . . . . .	67
4.7	Summary . . . . .	67
<b>Chapter 5 Conclusion and Future Work</b>		<b>69</b>
5.1	Conclusion . . . . .	69
5.2	Future Work . . . . .	71
<b>Bibliography</b>		<b>73</b>



# List of Tables

1.1	Different algorithms for medical image analysis. . . . .	6
2.1	FDA Qualified Biomarkers and Supporting Information [1] shared by Dr.Aamir. . . . .	10
2.2	Relates work for prostate histopathology image analysis and their results.	15
3.1	Details of data acquisition from different sources. . . . .	21
3.2	Summary of papers involving problems and solutions for histopathology image analysis. . . . .	22
3.3	Some Zernike moments physical explanation . . . . .	29
3.4	List of Zernike moment coefficients ( $Z_{pq}$ ) at each order p up to 5. . . .	30
4.1	Confusion matrix for cluster formed by endoscopy image dataset. . . .	57
4.2	Confusion matrix for cluster formed by lymph node image dataset. . . .	60
4.3	Confusion matrix for cluster formed by prostate tissue image dataset. .	62

# List of Figures

1.1	The set of steps involved in digital image technology used for decision making by processing and analyzing features of an image. [2, Figure 1]	2
1.2	Rise in the number of articles per year retrieved from Web of Science and Scopus using the keyword 'cell image identification'.[3]	3
1.3	Generic flow of the medical image analysis using artificial neural network	4
3.1	The pipeline diagram for the proposed method.	17
3.2	Different grades of cancer prevalent in prostate cancer.	20
3.3	Figure showing issues with H&E images like tear and use of marker (top), blurred image(bottom-left) and folded tissue section (bottom right).	24
3.4	RGB to HSV color conversion for gastrointestinal image(top), lymph node image(middle) and prostate tissue image (bottom).	25
3.5	Letter S, having similar moment, in left 3 blocks showing change in image scale and rotation. Letter K, in right, will have different moment value.	27
3.6	An image being mapped entirely within the unit circular disk known as Outer circle mapping.	28
3.7	The first 21 Zernike polynomials, ordered horizontally by azimuthal degree and vertically by radial degree	31
3.8	An interconnected group of circular nodes (artificial neuron) representing artificial neural network.	33
3.9	Various categories of machine learning with decision making approach.	36
3.10	Two-layer structure of the SOM training.	37

4.1	Healthy(on left) and Unhealthy(on right) images of endoscopy, lymph node tissue and prostate tissue in sequence from top to bottom. . . . .	42
4.2	System Implementation Design for Medical Image Analysis. . . . .	44
4.3	ZEISS ZEN lite software user interface. . . . .	46
4.4	CITU system cross-modal design architecture . . . . .	47
4.5	CITU system cross-modal design architecture. . . . .	48
4.6	The GUI for Neural Network Clustering app in MATLAB. . . . .	49
4.7	Represents the Topology structure of the Self-Organizing map grid, which traditionally is hexagonal or rectangular. . . . .	50
4.8	Graph showing z-score value values for z00 moment for healthy and bleeding endoscopy images. . . . .	52
4.9	A 4x4 SOM sample hit map for endoscopy image dataset. . . . .	53
4.10	A 4x4 SOM neighbor weight distance graph for endoscopy image dataset. . . . .	55
4.11	SOM weight planes for Endoscopy image data. . . . .	56
4.12	4x4 test data SOM sample hit map for endoscopy image dataset. . . . .	58
4.13	Represents a 7x7 (a) SOM sample hit and (b) SOM neighbor weight distance graph for the Lymph node image dataset. . . . .	59
4.14	(a) SOM weight planes for 12 inputs of lymph node image train data. (b) SOM sample hit for lymph node image test data . . . . .	60
4.15	Represents a 4x4 (a) Training data SOM sample hit, (b) Test data SOM sample hit, and (c) SOM neighbor weight distance graph for Prostate tissue image dataset. . . . .	61
4.16	Represents a 5x5 (a) Training data SOM sample hit, (b) Test data SOM sample hit, and (c) SOM neighbor weight distance graph for Prostate tissue image dataset. . . . .	63
4.17	SOM weight planes graph for 9 inputs of prostate tissue image train data. . . . .	64
5.1	A 3x3 Map showing clustering of healthy (marked with green) and sick (marked with red) of prostate tissue samples. . . . .	70

# Chapter 1

## Introduction

In this fast growing world of data, digital image technology is a rapidly emerging, multidisciplinary technology that can utilize the information from various fields to create a system, which has the potential to help humans in their field of work, for example in engineering, security, chemistry, biology, medicine, industrial automation, etc. The complete process of digital image technology can be categorized into the task of image acquisition, image digitization (which means the transformation of an image into digital form), image processing, and image analysis. An example of such a system could be ImageJ [4], which is an open source image processing program created for scientific multidimensional images. The advantages of such a system could be (a) we can infer large volumes of quantitative data in just a fraction of the time that manual image analysis would require, (b) it would reduce the human error giving consistent performance over time, (c) less human labor required making them free for more advanced work and (d) bring down the cost of analysis as multi-task will be possible. In regard to building such a system, Figure 1.1 shows various general steps involved in digital image technology. Any process requiring decision making based on qualitative or quantitative data extracted from images can use this technology to gain knowledge.

A similar decision making process is required when pathologists analyze tissue samples in glass slides or doctors look at the digital images of the endoscopy process, where they analyze various features of the sample medical image to diagnose the disease based on their years of experience of handling different types of samples. Overall, the medical image classification can be separated in the following three steps. First, gathering a

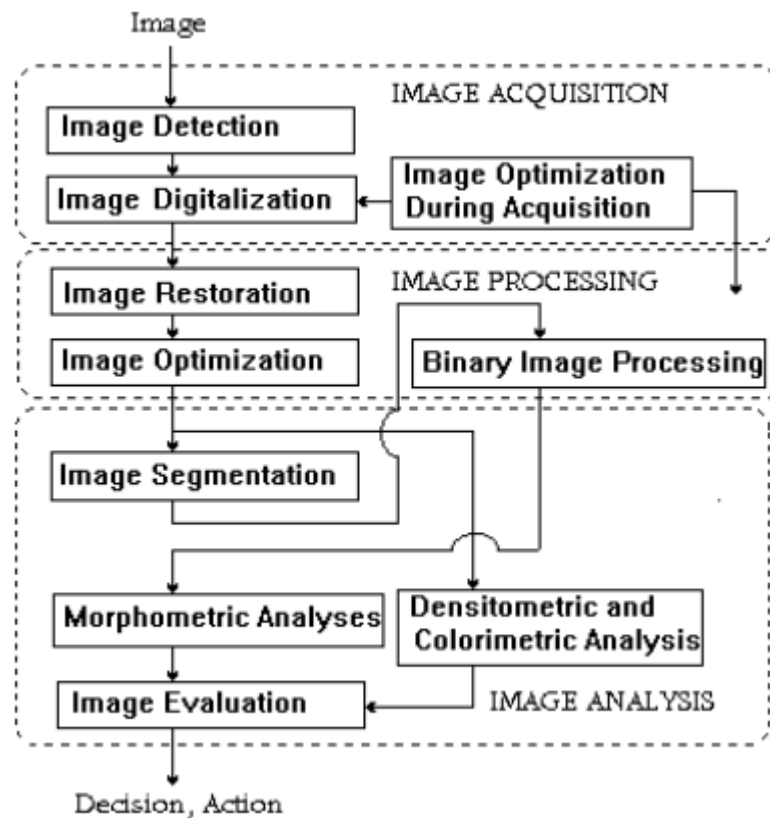


Figure 1.1: The set of steps involved in digital image technology used for decision making by processing and analyzing features of an image. [2, Figure 1]

good quality medical image. Second, extracting the relevant features from the images and third is building a model using these features to classify the dataset.

In the past few decades, the field of medical science has used image analysis benefiting significantly from the precise, fast, repeatable, and objective measurements made by computational resources. These quantitative measurements contribute to the analysis of structure and function in normal and abnormal cases by addressing many aspects of the data, such as tissue shape, size, texture, and density; musculoskeletal angle, kinematics, and stress; as well as ventricular motion, myocardial strain, and blood flow. The shape of tissue structures or organs is of particular interest in the visual interpretation of images, and automated techniques provide many quantitative measures that can contribute to the examination. The smoothness or homogeneity of the tissue is

also often used in visual examination to assess the state of the tissue. However, these are not the first steps in diagnosis, except in case of urgency or severe health concerns. Usually, a patient describes their symptoms based on their experience and discomfort to a medical doctor. Upon further analysis and considering the patient data and their family history a doctor does a second set of checks with stethoscope or blood pressure measurements. A more detailed diagnosis is gathered using maybe a blood or urine sample. Only when these reports raise a suspicion of a particular disease is when a doctor would suggest a medical image data which would further be examined to check whether the suspicion can be confirmed or excluded. These image data along with the detailed information collected during diagnosis is stored and recorded in some of the hospitals. This is mostly done to keep the records available for different purposes. But more efficient use of this data is when it is processed to find new techniques and innovations in the field of science. One such innovation is the image analysis. We can see the growing trend in the use of image analysis techniques by graph in Figure 1.2 showing the number of publications over the years in this field.

Due to this advancement in technology of image processing and storage, it has now

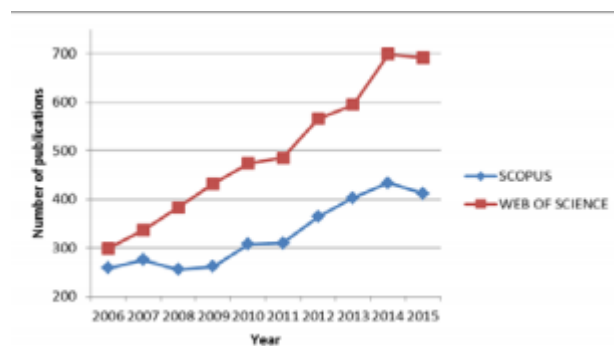


Figure 1.2: Rise in the number of articles per year retrieved from Web of Science and Scopus using the keyword 'cell image identification'.<sup>[3]</sup>

become easier and economical to have digital management of these medical samples, store detailed reports and pathology orders. So much so that now there is the digitization of histopathology slides and even the use of computer vision to observe these samples, which in the future aims to replace the current use of the optical microscope as the primary tool used by pathologists. Today medical systems produce a large number of digital images that contain a wealth of information. However, not all of this infor-



mation is relevant for the task at hand, most of the time the relevant data is hidden in the pixels along with a lot of additional noise. For instance, different medical images can be obtained from different sources having varied focusing regions, contrast, and white balance. Moreover, medical images have a different texture and pixel density based on their inner structure. Using a traditional set of features to classify medical images would lead to the inefficient characterization of certain classes [5]. Before deep architecture arrived, most of the studies have used models relying on the shape, color, and/or texture features as well as their combinations [6],[7]. These models have shown some good results. However, the biggest challenge that lies with these models is their lack of generalization due to the use of low-level features that fail to represent high-level problem domain concepts. In contrast, deep structured learning [8],[9] has shown huge success in a very short time for the non-medical image field. Although there still are studies using medical images with deep structure learning, they have shown relatively lesser success in providing a generic approach. Therefore, in this work, an artificial neural network (shown in Figure 1.3) is applied to find a generic approach of diagnosing confounded medical images using the high-level features extracted from the corresponding images.

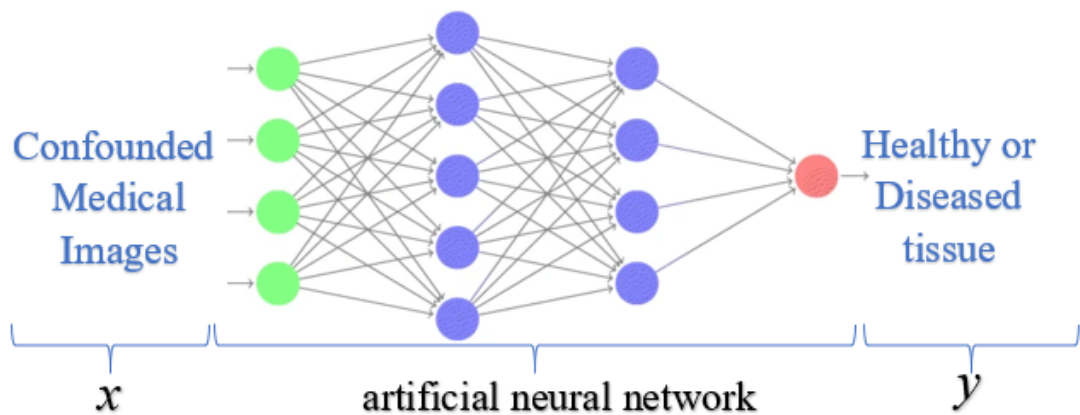


Figure 1.3: Generic flow of the medical image analysis using artificial neural network

## 1.1 Problem Definition

The previous section introduces the concept of digital image technology and how it can be used in the field of medical science. There are many challenges at each step starting from image acquisition (which could even decide the direction of research) to optimization, storage, and other related steps. Such challenges are addressed in this research and provide a basis for a good solution. However, looking at the big picture, the more important question would be how can we effectively distinguish between healthy and sick images?

Unlike the images of cats and dogs which have shown huge success in neural networks, the medical images are more complicated, involving various small parts that are moving and evolving with time. Another difference between them would be the requirement of a higher success rate in classification due to their dependency on saving human lives. Therefore, when it comes to classifying medical images there are two significant challenges that need to be addressed.

1. How can we extract effective features from a medical image dataset?
2. How to quickly and efficiently separate this dataset into related images?

Thinking of the feature extraction part, it seems logical to formulate the idea of segmenting the image into the region of interest, such as cell nuclei, glands, lobule formation, etc., that contains relevant information and create a large feature vector out of features like edges, corner, ridge, etc. This approach might work for one type of medical data but is not sufficient to cover different types of medical data or provide a generic solution. Table 1.1 provides a list of such implementation for different medical images. Using this technique would require the implementation of a broad range of image segmentation algorithms.

For medical images, the choice of a classifier is also a very important factor. Using these classifiers over the extracted features can efficiently separate the chosen dataset into related images. However, deciding the classifier to use can be challenging as each has its own advantages and shortcomings. As we discussed, deep learning has shown huge success in dealing with image and learning on its own. But this learning can be supervised, semi-supervised, and unsupervised. How to decide which one to choose

Table 1.1: Different algorithms for medical image analysis.

<b>Authors, Year of Publications</b>	<b>Organ</b>	<b>Method for Classification and Segmentation</b>
Elaine Yu et al. 2011	Breast	Color Gradient Active Contour, Hierarchical Normalized cut.[10]
Andrew Janowczyk et al. 2010	Prostate	Geodesic Active Contour.[11]
Hiroyuki Shimada et al. 2009	Follicular Lymphoma	Self Organizing Maps, Texture Classification using Non Linear color Quantization.[12]
Omar S. 2010	Meningioma Tumor	Texture Classification using fractal features.[13]
Melih Kandemir et al. 2010	Colon Glands	Segmentation using Object graph approach.[14]
Jyotirmoy Chatterjee et al. 2009	Oral Mucosa	SVM [15]
Metin Gurcan et al. 2011	Follicular Lymphoma	Color Texture cell Segmentation.[16]
Metin N. et al. 2011	Breast	Gaussian Mixture Model based segmentation [17]

from these? Well, in the majority of cases the medical images are unlabeled due to manual challenges and limited medical experts. Moreover, deciding labels will reduce the scalability the of target function. This logical decision making helps in narrowing the options and finding a rational solution.

## 1.2 Contributions

The key contribution of this work is the presentation and analysis of a broad medical image clustering approach which can be used in helping pathologist and doctors to diagnose different medical images. The most noticeable features of the classification of medical images approach taken in this research are as follows:

- The approach isn't limited to any specific image used in medical science. Rather, it can be expanded to more than the 3 different types of images taken in this research. This is possible by using a model that utilized a global set of features rather than using image specific segmentation techniques.
- Due to the techniques used, this solution is scalable in terms of both medical image representation and clustering requirements. This would allow the user to find an optimal solution based on the problem at hand, giving them more flexibility and control over the program.
- The results presented using this approach are easily understood and interpreted with help of simple visualization. As humans can easily grasp more complex information with simplicity using visualization. This would allow pathologist to make easier and faster decisions.

Apart from these core implementations of the approach taken, this research work has contributed in terms of knowledge required while analyzing some of these medical images. By using the adequate techniques to gather and represent knowledge, this research work contributes towards the identification of methods and their purpose for medical image analysis by taking inputs from medical experts in their respective fields

of research. Furthermore, this dissertation has possibility for publishing results in scientific and medical journals.

### **1.3 Structure of the Dissertation**

The remainder of the dissertation is divided into the following four chapters. Chapter 2 explains the motivation to pursue this research along with the literature review that covers the important concepts referred to in this dissertation, existing solutions, and the learning techniques used. The pipeline design for the chosen approach, knowledge and data acquisition details, and description of the algorithms used are explained in chapter 3. The implementation detailing the system architecture, dataset description, and the technologies used are elaborated in Chapter 4. This chapter also provides a set of case studies for evaluation of the approach, system design, and performance achieved when running under various scenarios. The limitation of this work, conclusion, and the scope for future work are later discussed in chapter 5. The details on conversations with medical experts are given in the appendix section of the dissertation.

# Chapter 2

## Motivation and Literature Review

The following chapter discusses the motivation (see section 2.1) behind the work introduced in Chapter 1. It also provides the literature review (see section 2.2) for the concepts used in this dissertation, a concise review of related work in this area, and a comprehensive review of the dataset and evaluation techniques used.

### 2.1 Motivation

The human brain is unparalleled when it comes to 2D image understanding and analysis. This is supported by the fact that half of the brain is devoted to the interpretation and processing of all visual data received through our eyes. However, what it's not good at is analyzing high dimensional images, like the ones used in the field of medicine. Our brains find it hard to in quantifying all the information available in a complex image and require training to do so. Although even if it could, it can't reconstruct images to enhance specific features. These are some general problems that a digital imaging system is really good at handling. If designed well, such a system takes the complex image part and reduce them into more manageable tasks that human can easily do while providing a standard to complete method.

Moreover, when trying to diagnose a disease like cancer one of the most common clinical use techniques involve biomarker testing. Biomarkers provide information about a disease beyond the standard clinical parameters using molecules for detection or evaluation [18]. Biomarkers can be DNA, RNA transcripts, metabolites, proteins,

Table 2.1: FDA Qualified Biomarkers and Supporting Information [1] shared by Dr.Aamir.

<b>Qualified Biomarker(s)</b>	<b>Abbreviated Biomarker Description</b>	<b>Abbreviated COU</b>
Albumin, $\beta$ 2-Microglobulin, Clusterin, Cystatin C, KIM-1, Total Protein, and Trefoil factor-3	Urinary nephrotoxicity biomarkers as assessed by immunoassays	Safety biomarker to be used with traditional indicators to indicate renal injury in rat
Clusterin, Renal Papillary Antigen (RPA-1)	Urinary nephrotoxicity biomarkers as assessed by immunoassays	Safety biomarker to be used with traditional indicators to indicate renal injury in rat
Cardiac troponins T (cTnT) and I (cTnI)	Serum/plasma cardiotoxicity biomarkers as assessed by immunoassay	Safety biomarker to indicate cardiotoxicity in rats, dogs or monkeys when testing known cardiotoxic drugs and may be used to help estimate non-toxic human dose
Galactomannan	Serum/bronchoalveolar lavage fluid biomarker: as assessed by immunoassay	Diagnostic biomarker used with other clinical and host factors to identify patients with Invasive Aspergillosis
Fibrinogen	Plasma biomarker as assessed by immunoassay	Prognostic biomarker used with other characteristics to enrich for COPD exacerbations
Total Kidney Volume (TKV)	TKV as assessed by MRI, CT and US	Prognostic biomarker with patient age and baseline glomerular filtration rate for Autosomal Dominant Polycystic Kidney Disease

or epigenetic modifications of DNA, etc. Table 2.1 provides a list of FDA approved biomarkers.

These biomarkers are effective, but the cost associated with them are high. The cost can vary from country to country and laboratory to laboratory, but overall, the measurement of environmental chemicals in the body could be very expensive. Sometimes there could be more than one biomarker involved in finding a better result, this increases the cost of the test even further. Compared to this the process of taking images of the tissue samples is much cheaper and even more cost effective in terms of storage. It also has a long-term benefit of being used as training or testing data to find new techniques of diagnosis using the field of computer vision.

Therefore, the cheaper and effective use of images along with a better analysis of these images using computers makes an effective argument for the need of a system that can process such information rich images and extract meaning out of them which can make the task of humans easier and faster.

## **2.2 Literature Review**

This section presents the details of studies performed in various techniques used in research showing key challenges faced and their advantages, which act as motivation for this work. The literature review has been categorized into subsections based on the similarity of studies and techniques being discussed.

### **2.2.1 Imaging in Histopathology**

Before reviewing the literature on the representation and analysis techniques, let's see an overview of the imaging process for histopathology images used in this paper. In a usual setup, tissue samples are sent to pathology labs for analysis after performing biopsies in operation rooms. In [19], the process of imaging from tissues is detailed. The first step includes formalin fixation and embedding in paraffin. Then a section of thickness from 3 to 5  $\mu\text{m}$  is sliced, using a high precision tool called microtome, from the paraffin blocks and are placed on glass slides. The nuclei and cytoplasm, area of interest in tissue, cannot usually be seen on the mounted sections. Hence, they required to be highlighted using dyeing with stains.



Hematoxylin-Eosin (H&E) staining is one of the most commonly used methods in this process [20][21], specifically in the analysis of tumor tissue microscopic images. Hematoxylin makes the nuclei blue/purple by binding to DNA and eosin dyes other parts (cytoplasm, stroma, etc.) to pink by binding to the protein. After this process, the slides are sent to pathologists for examination. Eventually, as digital pathology became common, the slide digitization was added as an extra stage to this workflow [22]. The initial process of taking still images using digital cameras attached to the microscope was replaced by a WSI scanner, which higher throughput at a lower cost. Also, they are capable of handling the entire scanning process automatically, including loading of slides, detection of relevant regions, taking images, storing, etc. Imaging is actually an important part of the complete process and any issues like out of focus or overlapping of parts can have a downside effect in the complete process of evaluation.

### 2.2.2 Image Feature Vectors

In [23], the author discusses different image features for building Content-based image retrieval (CBIR). It classifies features into low-level or high-level features. The low-level features are helpful in removing the sensory gap between the information in a description derived from an image and the object in the real world [24]. These could include features that reflect shape [24], color [25], texture [26], etc. While the high-level features are referred to as semantic features and use useful in removing the semantic gap between the interpretation of the same data and the information extracted from the visual data. However, due to visual data understanding inconsistency between different users, it's difficult to eliminate the semantic gap.

Using color moment invariants [27] implements an image retrieval system. In that, the color representations are calculated from each image rather than being limited in a given color space. This allows the feature set to be compact and accurate. Moreover, by taking advantage of a two-stage clustering method it adapts itself to the context of the image. In [28], the author presents an effective image similarity calculating approach by representing the image as a graph. In this, the image colors are first quantized and for each color, a histogram is created. Then a weighted undirected graph is created for different colors of the database image and query image. Finally, to find the distance between two images it sums all the colors and calculates minimum

cost matching for each graph. The method is both rotation and translation invariant making the technique quite robust.

### 2.2.3 Image representation using moments

In [29], the comparison between global and local features are mentioned when it comes to medical image analysis, where local features fail to capture the gross view of an image a global feature like Zernike moment is a more effective and robust solution for the biomedical image. Zhang and Lu in [30] describe various moments and their use. Their evaluation on the counter-based method and global region-based method shows how global features like moments, that extract statistical distribution of region pixels, make effective use of all the pixel information within the region. However, moments like geometric and complex moments can be sensitive towards noise and even contain redundant information [31]. This is mostly because of the reason that the kernel polynomials are not orthogonal. Zernike moment, an orthogonal moment, is less sensitive towards noise and has been used in various applications like prostate ultrasound [32], scanning brain tumor [33], cell image retrieval [34], etc.

### 2.2.4 Learning technique in medical images

There are many papers published in recent years that have used a neural network to learn a function for medical image analysis. In [35], to classify breast histology images into four classes- benign lesion, invasive carcinoma, in situ carcinoma, and normal tissue, the authors used a convoluted neural network(CNN) to extract a set of features from the image while training the model using support vector machines (SVM). In [36], the output of a mid-layer pre-trained network for prostate cancer is fed into random forest and SVM classifiers. A patch-wise accuracy of 81% and an image-wise accuracy of 89% is achieved using 10-fold cross validation. In [37], authors have used technique requiring a few labeled samples. In this work classification of colon histopathology images is performed using a multiple instance learning framework. The author compares the performance of a weakly labeled approach to a supervised learning approach with 94.52% and 93.56% accuracy respectively

## **2.3 Conclusion**

The above sections discuss literature about the various components that are part of this research to find the strength and weaknesses of work done in a similar field of study. Moreover, it provides an insight into the decision-making process while working with medical image processing. This work takes advantage of the literature discussed to get ideas that enable the chosen design and implementation. Table 2.2 is a brief summarization to evaluate the approach and results of similar studies using prostate histopathology images.

Table 2.2: Related work for prostate histopathology image analysis and their results.

Authors	Data	Classification Type	Evaluation	Results
Karimi et al. 2018 [38]	231 Patients, 333 cores & 1600 patches,	Gleason grade 3, 4 and 5	Patient-based cross-validation	Benign vs Cancer: accuracy 90.2%, Gleason grade 3 vs grade 4& 5: accuracy 76.6%.
Tomaszewski et al. 2010 [39]	20 Patients & 40 slides,	Benign vs Cancer	Slide-based cross-validation	Sensitivity 0.87 and specificity 0.90.
Madabhushi et al. 2012 [40]	58 Patients & 100 slides,	Benign vs Cancer	Slide-based and patient-based	Image accuracies 0.69, 0.7 0.87 and 0.69; Patient accuracies: 0.74, 0.66 and 0.57.
Gaed et al. 2013 [41]	15 Patients, 50 slides & 991 patches,	Benign vs Cancer; high-grade vs low-grade	Patch-based cross-validation	Benign vs Cancer: accuracy 0.90 and high-grade vs low-grade: accuracy 0.85.
Sarkar et al. 2014 [42]	29 Patients & 317 patches,	Benign, Gleason grade 3 & Gleason grade 4	Patch-based 10 fold cross-validation	Gleason 3 vs 4: 0.87.
Fricker et al. 2018 [43]	886 Patients,	Gleason grade 3, 4& 5	Patient-based	Recall on test set: 58%.

# Chapter 3

## Method

### 3.1 Introduction

There has been a significant advancement in image analysis over the last two decades, especially in the field of medical imaging and computerized medical analysis. These advancements have led researchers to look for more advanced ways for the diagnosis and treatment of various evolving diseases. At the same time, the success of machine learning algorithms at image recognition tasks provided a direction for exploring new ways of analyzing these electronic medical records and diagnostic imaging.

This overlapping success of machine learning algorithms in computer vision with that of medical image digitization has opened up opportunities in all fields of medicine, starting from the discovery of drugs to clinical diagnosing, potentially changing the way the current medical system works. We can see this trend in electronic health records (EHR) of US that grew 4 times between 2007 to 2012 for office-based physicians [17]. Similarly, there are transformations in pathology labs to turn towards a completely digital workflow [22]. In addition to digitally storing these tissue samples, which requires digitization of histopathology slides and using computer monitors for analysis, pathologists are now replacing the optical microscope as the primary tool. Furthermore, due to the prevalent nature of these disease, there is a large number of samples being analyzed each day by the pathologist, which sometimes can be tedious and are affected by observer variability [44], [45]. The change of digitizing medical images has more benefits than just advantages in speed and objectiveness compared to

manual approaches. One such benefit compared to glass slides or other physical samples is that these digital images allows the possibility of doing quantitative automatic image analysis using various available techniques and can utilize image enhancement techniques that are not possible with normal slides.

Figure 3.1 shows the steps performed in the ‘Medical Image Analysis’ algorithm proposed in this research and explained in the later section of this chapter.

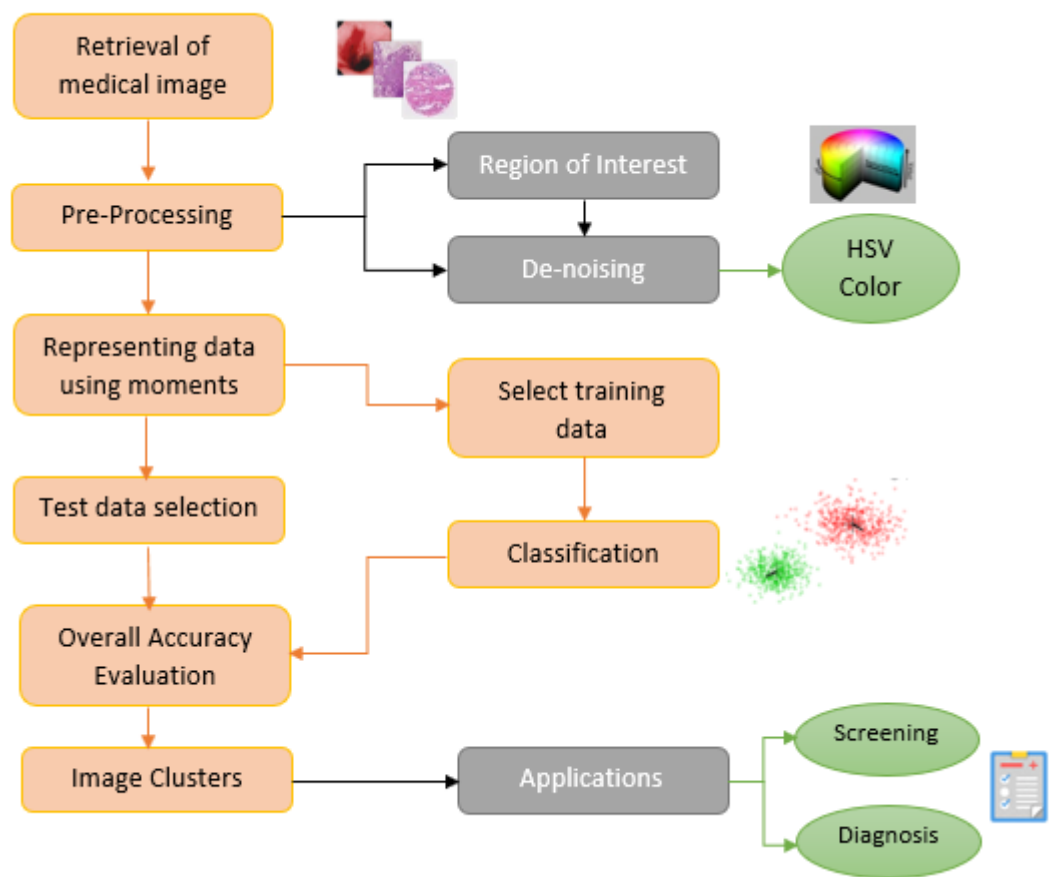


Figure 3.1: The pipeline diagram for the proposed method.

The input image data is explained in section 3.2 along with the different types of images and their source. Pre-processing techniques used in this implementation are explained in section 3.3 following details of the data representation technique in section

3.4. Section 3.5 will talk about the classification algorithm used. Finally, ending this chapter with a brief conclusion in section 3.6 and the rest implementation details will be described in the next chapters.

## 3.2 Knowledge and Data Acquisition

While working with medical images it matters what the source of the image is? What is the purpose this image was generated for? And lastly, gathering knowledge on what are the important features in the image being used? To tackle all these challenges this research refers to various sources of medical images, from open source to gathering images from medical institutes, to analyze the success of the algorithm and talk with medical experts to validate the approach taken.

### 3.2.1 Knowledge Acquisition

Kim and Courtney define knowledge acquisition as "the process of gathering knowledge about a domain, usually from an expert, and incorporating it into a computer program" [46]. Although it's possible to gather this knowledge from various other sources like textbooks, articles, and journals. It often considered a bottleneck in the process of development due to the time and complexity involved in the process. The complexity in this work is mostly due to the involvement of an interdisciplinary field that requires expert advice and provide a more reliable result.

Hence, for this research, we consulted experts to gain knowledge about data and understanding the challenges involved in processing such information by a human mind. There are various existing approaches for knowledge acquisition like commentary, interviews, observation, etc. Fortunately, to support this research two medical experts agreed to provide their invaluable time to help understand the analysis technique and provide medical images discussed in the later sections.

Dr. Aamir Ahmed, who is the head of the stem cell and prostate cancer group at King's College London, has been working with a wide range of molecular biological, histochemical, biochemical, live cell imaging and high throughput (genomic, proteomic, electrophysiological, and tissue imaging) techniques to address fundamental questions regarding Wnt signaling and how this knowledge could be translated into better thera-

pies and quantitative biomarkers of cancer. He has written several papers, [47][48][49], as part of his research in biomarkers for prostate cancer diagnosis. In meetings with Dr. Aamir we discussed a few of the fundamentals of Hematoxylin and eosin stain(H&E) prostate cancer tissue, which is one of the principal stains used in histology [50][51][52], and how the diagnostics analyze these images under a microscope. Furthermore, Dr. Aamir has also provided H&E stained prostate tissue samples that have been used in this research.

Rui Henrique, Professor at the Department of Pathology and Molecular Immunology of ICBAS-UP and appointed Chairman of the Board of Directors of IPO Porto, has years of experience and expertise in research and diagnostics focused in Hematopathology and Uro pathology, and is devoted to the understanding of the role of epigenetic alterations in tumorigenesis, as well as the development of novel cancer biomarkers based on the epigenetic and genetic characterization of tumors. Rui has published several papers, [53][54], related to prostate cancer and use of biomarkers. During the interview with Rui, he explained the structure of prostate tissue, color due to Hematoxylin and eosin staining, Gleason score, and few other challenges faced by a pathologist. Below is a brief summary of overall interviews conducted.

- Analysis of H&E tissues by pathologists is mostly intuitive due to years of practice of looking at H&E images. However, there are key structures to look for. Such as:
  - Small glands in front of large glands.
  - Haphazard distribution.
  - The higher density of nuclei and larger nuclei (nucleomegaly).
  - Presence of prominent nucleoli.
- Pathologists usually use 40x zoom under a microscope and sometimes require up to 400x to be certain.
- Sometimes (~5%) Immunohistochemistry (IHC) is used in diagnosis to confirm borderline cases. However, it's not always accurate. It works due to the presence (or absence) of basal cells, detected by specific antibodies against it combined with racemase expression in luminal epithelial cells.



- Gleason score (shown in figure 3.2) is based on how much cancer looks similar to small benign glands when viewed under a microscope. Based on how the cancer cells are arranged a score on a scale of 3 to 5 is assigned (patterns); due to morphological heterogeneity, in many cases more than one pattern is present. Depending on the proportion, predominance and potential aggressiveness, as well as whether grading is performed on prostate biopsy or surgical specimen, the final grade combines two figures: e.g. 3+4=7 (grade group 2); if only one pattern present, then, the pattern is doubled: e.g., 3+3=6 (grade group 1).
- In H&E staining Hematoxylin colors the nuclei of cells blue or dark-purple and Eosin stains the cytoplasm and some other structures including extracellular matrix such as collagen in up to five shades of pink. In cancer glands there is usually a blue-tinged secretion in the lumen and, sometimes, pink crystal-like structures (crystalloids).

Find the questionnaire and minutes of meeting attached at the end of dissertation.

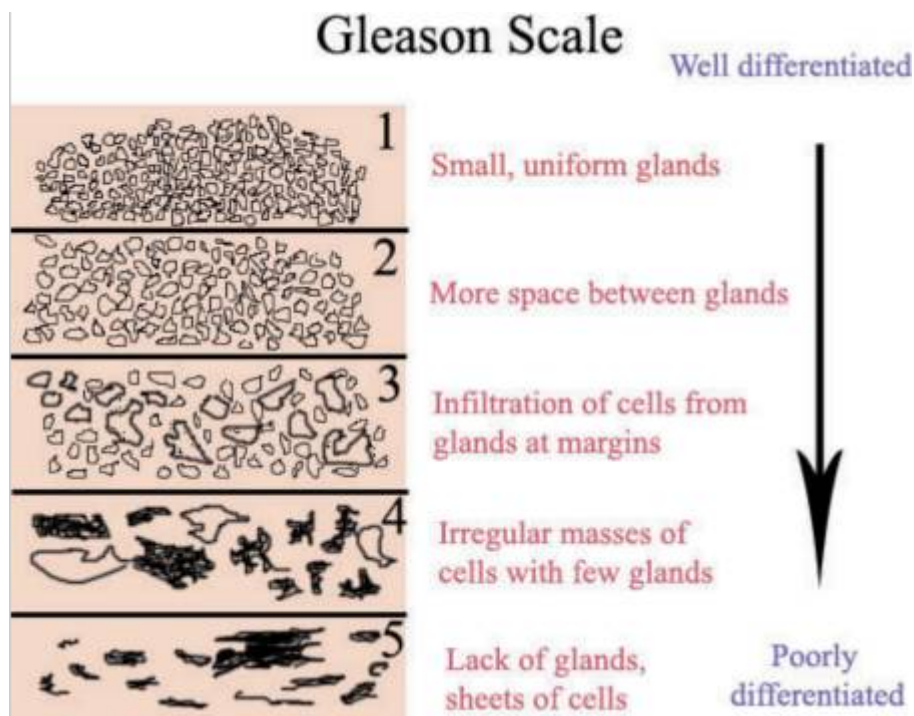


Figure 3.2: Different grades of cancer prevalent in prostate cancer.

### 3.2.2 Data Acquisition

As discussed, there has been a boom in medical image analysis. However, it has certain challenges when it comes to the acquisition of such medical data. For instance, in just the last few years the size of medical image data has gone from Kilobytes to Terabytes. Even though this means that we have a better quality of image data, it comes with a significant cost in processing and storing this data. Another major challenge that comes with using human medical imaging data and their corresponding meta-information is the concern with privacy and ethical aspects of data access. Fortunately, in this research no human tissues were used, only anonymized images were taken with just healthy or sick image labels. Hence not requiring any ethics approval for their analysis.

Table 3.1: Details of data acquisition from different sources.


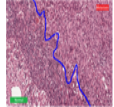
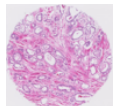
Field	Source	Type	Purpose	Sample
Endoscopy	Medical Journal	Gastrointestinal image	Pedagogical	
Breast Cancer	Open source dataset	H&E stained whole slide images of lymph node sections	Digital pathology	
Prostate Cancer	Kings College Prostate Cancer Research Centre	H&E stained whole slide images of prostate tissue	Digital pathology	

Table 3.1 shows the details of medical image data used in this research in order of research progression. It includes data from 3 different sources, types, and fields of study, along with their intended purpose for creation. This variability in data shows how the approach taken in this research is generic and is applicable to various kinds of medical image data. However, it must be noted that the intended purpose for endoscopy images used was pedagogical as it was published in medical journals from where it was scanned and used in this analysis. Such images essentially are low pixel density. Although it provides analysis results, its not the gold standard for medical image analysis. In contrast, the rest two categories of images are taken by experts in labs and are considered the gold standard for medical image analysis in their respective

fields. The prostate tissue images used were taken using slides at 40x magnification and with the help of a high-definition scanner, i.e. Nanozoomer slide scanner (Hamamatsu Photonics UK Ltd, Welwyn Garden City, UK) [55], the image are converted into a digital format. This high definition image capture technology helps to digitally give a resolution of up to 60X, hence the high pixel density and large file size.

These images are later used for classification by implementing machine learning techniques. But there are many major challenges that are faced when gathering and handling such medical images. Table 3.2 shows various papers that present solutions to known challenges when dealing with histopathology medical images which are applicable to other source of medical images too.

Table 3.2: Summary of papers involving problems and solutions for histopathology image analysis.

Problem	Solution	Reference
Very Large Images	Object level classification or case level classification summarizing patch	Bag of Words of local structure [56], Markov Random Field [57] and random forest [58]
Insufficient labeled images	Active learning	Hypothesis space reduction [59], Uncertainly sampling [60], and variance reduction [61]
	Multiple instance learning	deep weak supervision [62] and Boosting-based [63]
	Semi-supervised learning	Manifold learning [64]
Variation in magnification levels	Multiscale analysis	Texture features [65] and CNN [66]
Color variation and artifacts	Removal of color variation effect	Color augmentation [67] and color normalization [68]
	Artifact detection	Tissue-folds [69]

The table 3.2 gives a glimpse of the kind of challenges that researchers come by

when dealing with these images. These challenges don't just cause problems in the input image but also impact the complete end to end implementation. Problems like dealing with storage and processing capabilities to handle large images, find the right set of learning techniques for the unlabeled data, and sometimes even dealing with insufficient labels. Therefore, even though the medical image data has been increasing and can now be used in digital analysis, there still is some room for improvement for problems that are being tracked and new improved solutions are being implemented with time.

### 3.3 Preprocessing of Images

Pre-processing these images is an important step in complete medical image analysis. It aims to remove or suppress the unwanted distortions in images and to enhance some of the relevant features that are crucial for further processing.

All these medical images mentioned in the last section are taken using various different approaches. For instance, endoscopy images are taken using endoscopes that are swallowed by patients and includes a tiny camera. This camera transmits several images to the receiver over the course of the journey through the digestive tract. The flashlight, present in the camera, is used to show reflection in presence of moisture on the surface of the digestive tract. These reflections are hard to detect as their shape and color change in different images. Therefore, a minor change in external lighting (such as pale shadows, etc.) can cause a huge variation in RGB values thus changing the results significantly. To overcome this, Hue-Saturation value/Intensity (HSV/HSI) color coding is more informative in the detection of such specular reflection [70].

In the case of H&E scanned images, there can be issues that are completely unrelated to the actual tissue. For instance, there could be wrinkles in images due to the bending of tissue slices when placed on the slide; some segments could be blurred due to different thicknesses of some tissue regions; and sometimes there could be marking on images (Figure 3.3). These issues are present in the current chosen dataset as well. Due to such abnormalities, that can adversely affect the performance of the algorithm, we can apply techniques like color augmentation to provide better contrast between

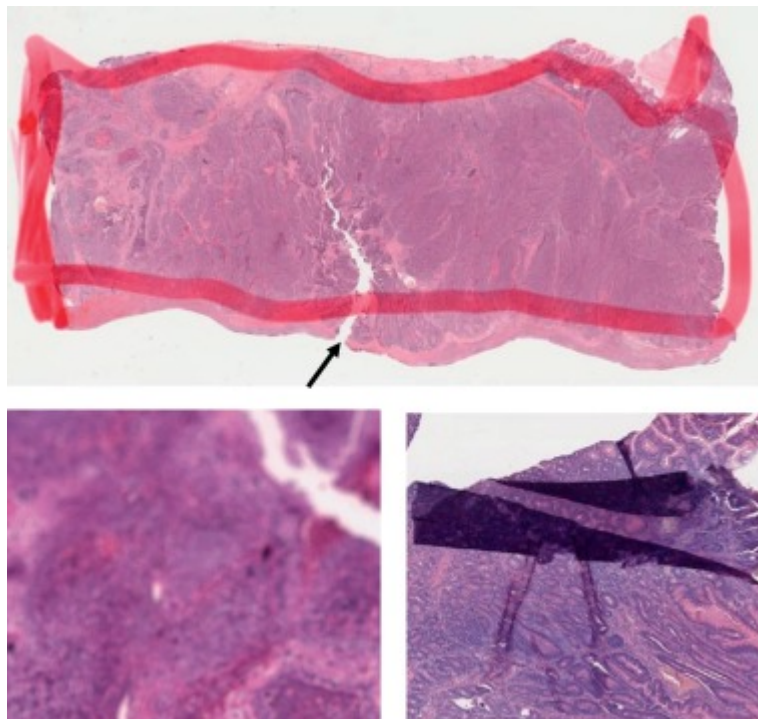


Figure 3.3: Figure showing issues with H&E images like tear and use of marker (top), blurred image(bottom-left) and folded tissue section (bottom right).

various regions using substantial color distribution provided by HSV color coding.

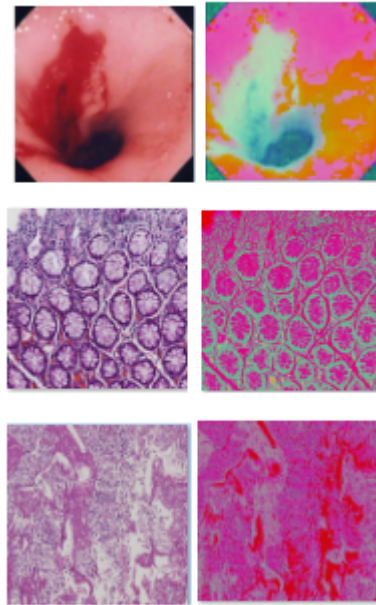


Figure 3.4: RGB to HSV color conversion for gastrointestinal image(top), lymph node image(middle) and prostate tissue image (bottom).

Furthermore, many of the common enhancement techniques based on contrasting are limited by the setting of parameters and application-based specifications. In color image processing, color distortions occur frequently due to the high correlation between the primary three colors, namely red, blue, and green. So, to avoid these distortions many have tried to convert this RGB image to an image with separate chrominance and luminance color space. Blotta et al. [71] studied the enhancement in RGB color images captured using an electron microscope and endoscope to Hue-Saturation-Intensity (HSI) image. The findings point the advantage in using HSV over RGB in terms of separating the intensity of the intrinsic color information, which is important when there is a change in lighting over the image taken. Figure 3.4 shows a sample of each image type when RGB to HSV conversion is applied.

## 3.4 Represent Image Using Moments

Medical images contain a large number of features such as color, texture, shape, and spatial layout. These features need to be represented in form of numerical attributes that can be extracted from images for purpose of representation and classification. However, considering a large amount of medical image data the conventional analysis list might not be long. For example, the Histopathology images used in this research vary from hundred to thousands of megabytes. Even if we consider a large storage system, it would be a partial answer to a larger problem. The available technique of using wavelets for image analysis has a known weakness for representation and detection of the contours of image objects. To overcome such challenges, new ways of image representation have been proposed which are based on theory of moments. Moments are already being utilized in medical image analysis since they provide a generic representation of any object, with both complex and simple shapes, that can be segmented in an image. Moments calculated from images can have more geometric and intuitive meanings compared to the low-level features like color, shape, or texture.

### 3.4.1 Image Moments

In the field of computer vision, an image moment can be defined as the weighted average of the image pixel intensities. Moments were first introduced by Hu in 1962 [72] and have since been used in various applications such as pattern matching, image normalization, contour shape estimation [73]. These moments comprise of nonlinear functions which are scale, translation, and rotation invariant. Meaning that the letter S in each block of figure 3.5 will give similar moment values. However, letter K will have a very different moment value from the rest of them. They also capture global information about the image and therefore do not require close boundaries like the Fourier descriptors.

Regular moments can be defined in form of projection of  $f(x, y)$  function onto the monomial  $x^p y^q$ . Here is the function to define regular moments.

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



Figure 3.5: Letter S, having similar moment, in left 3 blocks showing change in image scale and rotation. Letter K, in right, will have different moment value.

Where  $m_{pq}$  is the  $(p + q)^{th}$  order moment of the continuous image function  $f(x, y)$ . In the case of digital images, their integrals are changed to summations making  $m_{pq}$  as shown in equation 3.1.

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (3.2)$$

However, the regular moments are not orthogonal. The orthogonality makes the reconstruction of the original image from moments simplified and computationally easier. It also means that any two different polynomials in the sequence will be orthogonal to each other. Therefore, an absence of orthogonality implies the information content might have a certain level of redundancy.

To overcome this challenge associated with regular moments Teague [74] has recommended the use of orthogonal moments, which are based on the theory of orthogonal polynomials, like Zernike moments. Although there are other orthogonal moments available, the advantage of Zernike moments comes from its useful rotation invariant property. Meaning that rotating the image does not change value of Zernike moments and can be really helpful when applying to medical images such as that of gastrointestinal images.

### 3.4.2 Zernike Moments

Zernike moment is a kind of orthogonal complex moments and its kernel is a set of Zernike complete orthogonal polynomials defined over the interior of unit disc in the polar coordinates space [75]. These moments are named after the optical physicist



Frits Zernike, also awarded Nobel prize in 1953 for the invention of phase-contrast microscopy.

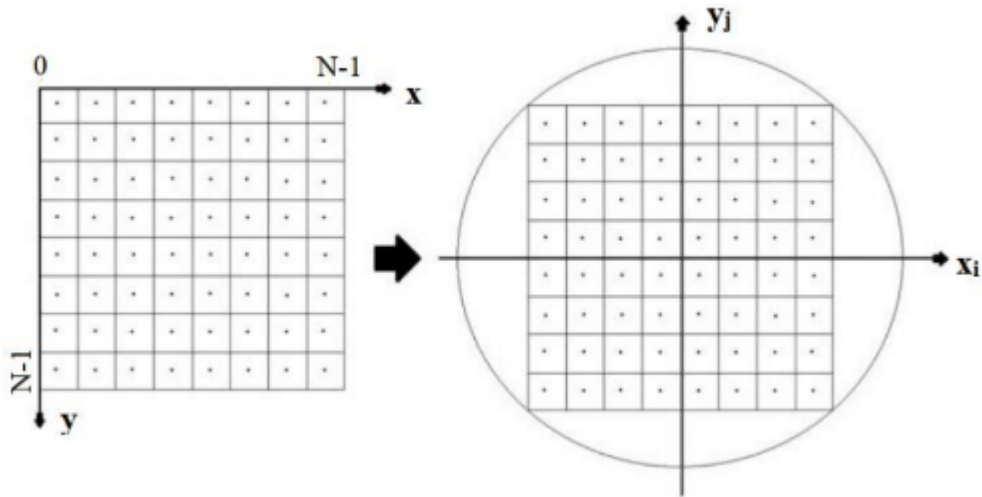


Figure 3.6: An image being mapped entirely within the unit circular disk known as Outer circle mapping.

The following function defines Zernike moment of order  $n$  with repetition  $m$  for a continuous image function  $f(x, y)$  over the interior of the unit circle, i.e.  $x^2 + y^2 = 1$ , as shown in equation 3.3.

$$A_{nm} = \frac{n+1}{\pi} \int \int_{x^2+y^2 \leq 1} V_{nm}^*(\rho, \theta) dx dy \quad (3.3)$$

where

$n$  Positive integer or zero.

$m$  Positive and negative integers subject to constraints  $n - |m|$  even,  $|m| \leq n$ .

$\rho$  Length of the vector from the origin to  $(x, y)$  pixel.

$\theta$  The angle between  $x$ -axis and the vector  $\rho$  in a counterclockwise direction.

For digital images, their integrals are changed to summations making  $A_{nm}$  as shown

in equation 3.4.

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x, y) V_{nm}^*(\rho, \theta), x^2 + y^2 \leq 1 \quad (3.4)$$

Zernike moment for a given image is computed by taking the center of the image as the origin and mapping the pixel coordinates to the range of unit circle, i.e.  $x^2 + y^2 \leq 1$  (seen in figure 3.6). Pixels that fall outside the unit circle are not considered in computation pointing to the known limitation of using Zernike moments.

Table 3.3: Some Zernike moments physical explanation

Moment	Explanation
Z00	Area
z11	xy-tilt
z20	Defocus
z22	Astigmatism
z31	Coma
z33	Trefoil

Medical Images such as histopathology or endoscopy images contain important structures, textures, and shape features. Zernike moments are very good at capturing such features as shown by Zhang and Lu [76]. This is mostly due to the orthogonal property of Zernike moments which allows it to have almost zero redundancy in the computed set of Zernike moment coefficients (seen in figure 3.7). Therefore, these moment values represent unique and independent characteristics of an image at different orders. They also describe a global feature set as they treat an image as a whole and gather the statistical distribution of the pixel intensities of the image, which is less

likely to be impacted by the changes in shape. As seen from equation 3.4, Zernike moments are computed using the summation process, making any noise introduced in the magnitude of Zernike moment coefficients negligible.

Table 3.4: List of Zernike moment coefficients ( $Z_{pq}$ ) at each order p up to 5.

Order p	Repetition q	p-q	$Z_{p,q}$
0	0	0(even)	$Z_{0,0}$
1	1	0(even)	$Z_{1,1}$
2	0	2(even)	$Z_{2,0}$
	1	1(odd)	-
	2	0(even)	$Z_{2,2}$
3	0	3(odd)	-
	1	2(even)	$Z_{3,1}$
	2	1(odd)	-
4	3	0(even)	$Z_{3,3}$
	0	4(even)	$Z_{4,0}$
	1	3(odd)	-
	2	2(even)	$Z_{4,2}$
5	3	1(odd)	-
	4	0(even)	$Z_{4,4}$
	0	5(odd)	-
	1	4(even)	$Z_{5,1}$
	2	3(odd)	-
	3	2(even)	$Z_{5,3}$
	4	1(odd)	-
5	0(even)	$Z_{5,5}$	

Moreover, these moments are both scale and rotation invariant. Meaning, if the analysis image is the rotated, up-scaled, or down-scaled version of an already analyzed image then using Zernike moment would allow us to accurately identify and match those images. Table 3.3 shows the physical meaning of some of these moments.

In this research approach, only low order Zernike moments have been calculated for analysis. The reason being the low order Zernike moments are more efficient in describing the complete information content in an image, unlike the high order moments that are more vulnerable towards noise [77][78]. Similar results have been seen in other applications that have used Zernike moments such as face recognition and image-based

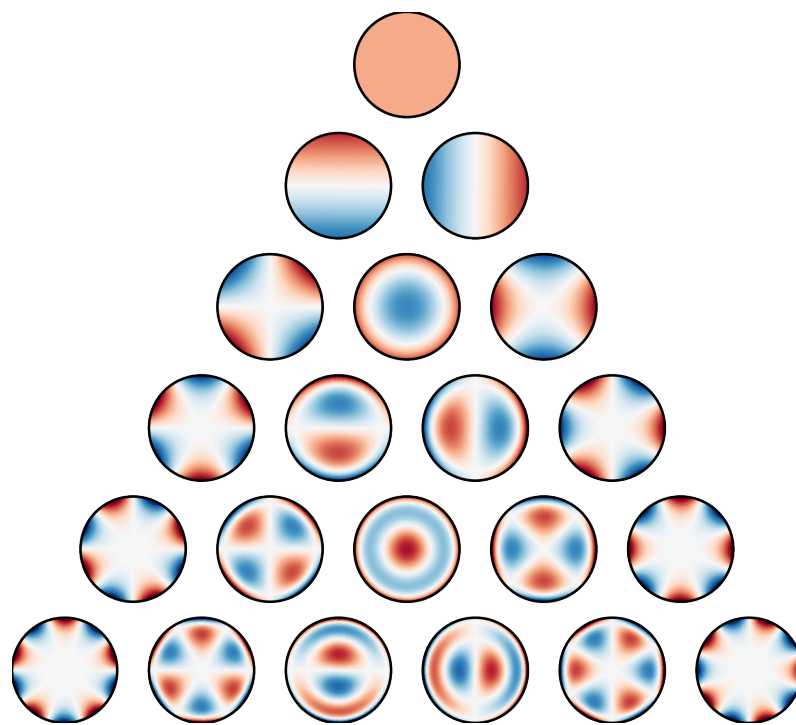


Figure 3.7: The first 21 Zernike polynomials, ordered horizontally by azimuthal degree and vertically by radial degree

retrieval system [79][80], pattern matching [78], etc. In [78], authors took images from noisy as well as noise free English alphabet characters to check the efficiency of noise resilience property of Zernike moments. In [79][80], authors have used an image-based retrieval system to compare the performance of many local and global descriptors (including Zernike moments). In these applications, authors have shown that using low order Zernike moment coefficients there is improved performance on the noisy image as compared to the noise free image. Thus, in this research of medical image analysis, we have used up to 12 low order Zernike moment coefficients, which corresponds to a maximum of 5 orders of moments for retrieval of histopathology and endoscopy images. Table 3.4 shows the list of Zernike moments calculated at each order  $p$  (up to 5) and repetition  $q$ .

The ‘-’ sign in the above table suggests that Zernike moment coefficient  $z_{pq}$  does not exist for the corresponding value of order( $p$ ) and repetition( $q$ ) whose difference, which is  $p-q$ , is an odd number.

Overall, Zernike moments, due to its various advantages like compact features, good accuracy, low computational cost, and robust feature retrieval, has been successfully used in various applications and shows an advantage in representing texture and shape features of these medical images.

### 3.5 Classification and Evaluation Methods

Diagnosing medical images is one of the main challenges of pathologists, with steps involving both acquiring quality images and good image diagnosis. For analyzing these images humans are limited to the number of images/ data that they can process, physical quality of the original image, presence of noise, incomplete visual search pattern and complex disease state. Therefore, various Computer Aided Diagnosis (CAD) systems have been commercialized over the last decade and have since been in clinical use. However, there is no fit for all systems available. Each of these computerized image analysis systems requires changes specific to the job and the imaging modality. Once the analysis is complete the results are shared with pathologists, which helps them in their task decision making while acting as assistance or a second opinion.

Machine learning (ML) over the years have shown some promising results in achieving similar tasks. The machine learning approach is based on using a set of methods that

can automatically find the pattern in the information given and later utilize this discovered pattern to predict future information or help in the process of decision making under various conditions. The most interesting aspect of machine learning is its data driven approach and figuring out results without much human intervention. Artificial Neural Network (ANN), shown in figure 3.8, is a part of machine learning that can resemble the multilayered human cognition system has proven ability in pattern recognition that can be utilized in such medical images. These systems are designed after the biological brain and are great at learning to perform new tasks by considering example data without being specially programmed to do such tasks.

## Artificial Neural Network

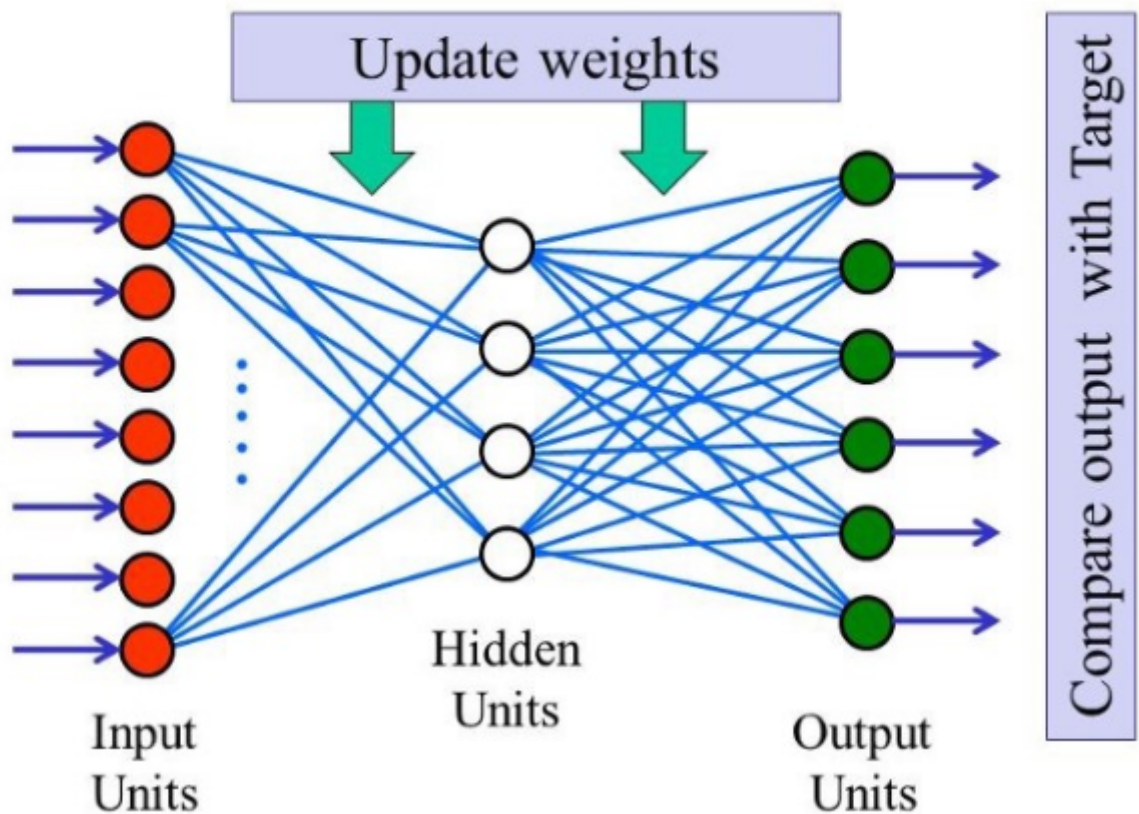


Figure 3.8: An interconnected group of circular nodes (artificial neuron) representing artificial neural network.

Although the Artificial Neural Network was introduced back in 1950, it hasn't gained popularity until recently due to various challenges faced like lack of computational power, absence of enough data to train the system, overfitting problems, and vanishing gradients. However, with advancements such as higher computing power after the introduction of graphical processing units (GPU) and the evolution of big data, these challenges no more hinder the unlimited possibilities offered by ANN. Advantages of ANNs, including Self Organizing Maps, are as follows:

- It can be used where other approaches must give up.
- They are easier to apply.
- Provides good evidence of the certainty due to the use of a test set.
- It is also applicable on heterogeneous data set.
- Data set can be used in a near-optimum way with ANNs.

ANN's also have a few disadvantages that can be summarized in the following points:

- Unless the algorithms are developed or a hybrid between ANN and a normal model is applied, there is no causality.
- The accuracy of predictions provided are sometimes limited in this approach.

The sudden rise in the use of machine learning and artificial intelligence requires medical experts to have an understanding of the technology, in order to know the abilities of these technologies and the impact and influence it will have in the future of the medical field. It's clear that in near future the adoption of such a system will happen, but in no means, it is a replacement of our medical experts. Rather it will be it will complement and enhance the work that our medical experts do. In this section, we will discuss details of the machine learning approach taken and also mention the reasoning behind the approach chosen. Finally, we also need to evaluate the machine learning algorithm to show the performance of the approach taken and can also be used to compare the results with any other implementation or system that is similar to this. The confusion matrix is used to evaluate the results and the details are discussed below.

### 3.5.1 Learning Approach for Medical Image Analysis

In order to help medical experts in their process of diagnosing the medical images acquired through the process of digitization, we need to choose a technique that can evaluate the important features of the image and classify them into different categories based on the requirement. Both the image features and the classification techniques matter here, as changing either would vary the results. For training such algorithms, the machine learning method can be categorized into a supervised learning approach and an unsupervised learning approach. The supervised machine learning approach uses some function that reproduces output by inferring from labeled training data. The data passed to this function is a numerical or nominal vector that contains the properties of input training data and their corresponding output values. If there is a continuous value in the chosen output data, then the process is categorized as regression. But if this output value is of categorical form then the process is categorized as classification. In contrast to the supervised learning approach, the unsupervised machine learning approach isn't based on the value of output data, rather it infers some method that can capture the hidden structure of the unlabeled input data. As the data passed to unsupervised learning is unlabeled, there usually is no objective evaluation of such a method. However, there are still some techniques that can be used to infer the performance of the system. Unsupervised learning is also considered similar to clustering in statistics, even though it encompasses various different solutions relating to summarization and key feature explanation, and points to the manner that relates to vector space representing the hidden structure, including clustering and dimensionality reduction.

Both supervised and unsupervised learning have shown promising results in the field of image analysis and medical imaging. However, in this approach, we have considered using unsupervised learning, which unlike the supervised approach is unbiased on how data is being analyzed and requires no manual effort of pathologists to create a large amount of label for the algorithm to work.

Based on the input images the pattern can be captured using simple two or three dimensional visualizations such as distribution curves, surface plots, and scatter plots [81][82][83]. However, if these descriptors are large in number then it becomes really hard to visualize these features. Again, the unsupervised clustering approach can be



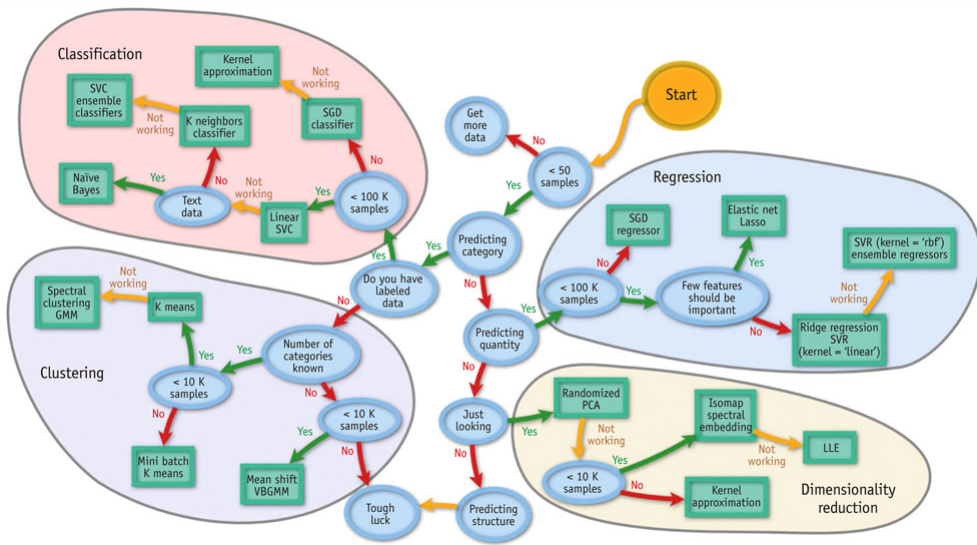


Figure 3.9: Various categories of machine learning with decision making approach.

useful here for reducing the features before visualization. Self-organizing maps, Hierarchical clustering, and k-means are some of the common techniques that are used in medical imaging to achieve such visualized clustering. Self-organizing maps have been employed in various techniques like patient stratification[84], interpretation[85], visualization and segmentation [82][86][87] in systems using medical images. Similarly, Hierarchical clustering is used in visualization and patient stratification [88][89]. K-Means on the other hand is mostly required for image and color classification, as well as visualization as part of the bag-of-features representation. For this research, we have chosen self-organizing maps for feature reduction and visualization over a 2-dimensional map.

### 3.5.2 Self Organizing Map

The Self-Organizing Map(SOM) is a clustering algorithm which is from the field of Artificial Neural Networks [90] and is trained using an unsupervised learning approach to reduce high n-dimensionality features into low dimensional (usually two-dimensional), discretized visualization of the input space of the training samples, known as a map. Also known as Kohonen map or network, the artificial neural network introduced by a Finnish professor Teuvo Kohonen in the 1980s. Although the dimensionality of the

space is reduced, the SOM is subject to a topological ordering constraint, thereby retaining the underlying structure of the input space [91].

Figure 3.10 shows an example RGB color coded image mapped over the output layer

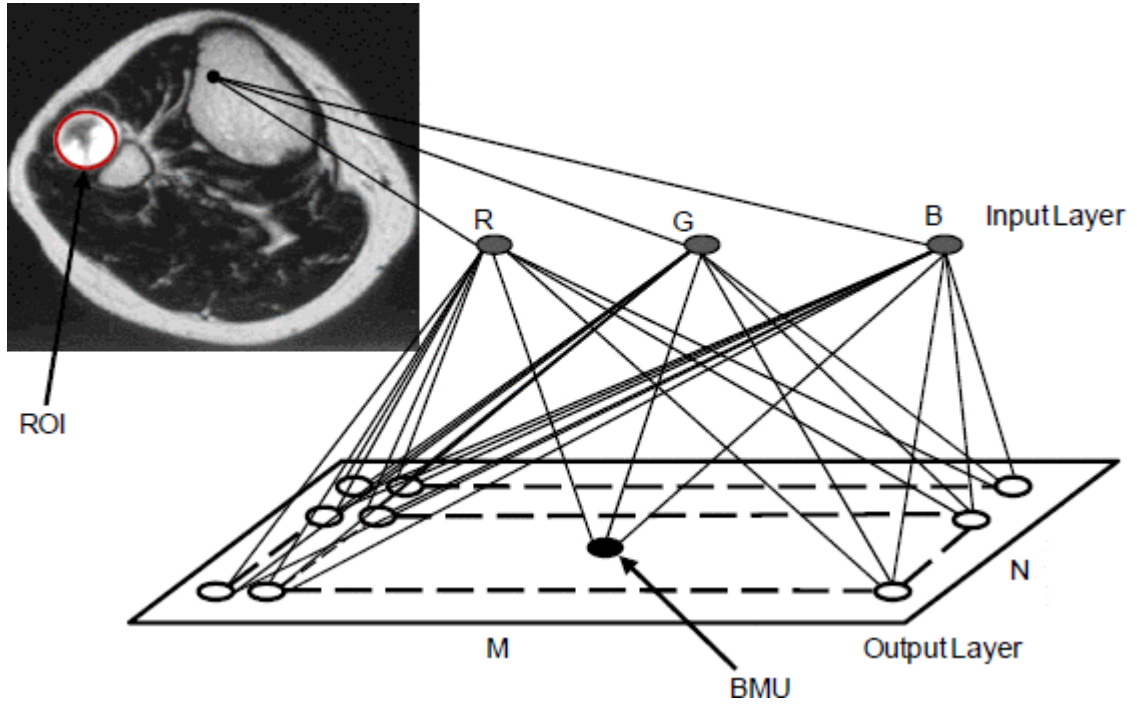


Figure 3.10: Two-layer structure of the SOM training.

using red, blue and green values of the image as the input vector. However, any other color model such as HSV or CMY could be used. Not just color but other relevant features that can be useful in representing the data that is being analyzed can be used in form of input vector.

$W_i = [w_{i1}, w_{i2}, w_{i3}]^T$  denotes the weighted average associated with the neurons in the output layer. Their values are randomly assigned initially where  $0 \leq i \leq M \leq N$  (i.e., the size of the output layer). SOM works iteratively, where in each iteration the best matching unit (BMU) or the neuron with the most similar weight vector can be found for an input vector. Additionally, the neighboring unit of the BMU are involved in a training process by pulling them close to the input vector. This is done using a

function that can be described by an exponential decay on time  $t$  as follows.

$$\sigma(t) = \sigma_0 \exp\left(\frac{-t}{\eta}\right) \quad (3.5)$$

where  $\sigma_0$  is the initial value of neighboring length and  $\eta$  is the time constant. Furthermore, as described by a Gaussian function  $\delta(t)$ , as we move away from BMU to the edge of the whole neighborhood the effect of learning gradually decreases. Also, as the learning process converges, the learning rate  $L(t)$  also decreases over time. Consequently, the weight vector of every neuron in the neighborhood of the BMU gets adjusted as per equation 3.6.

$$W_i(t+1) = W_i(t) + \delta_i(t)L(t)(V(t) - W_i(t)), \text{ if } \|i - BMU\| \leq \sigma(t) \quad (3.6)$$

where  $\delta_i(t) = \exp\left(-\frac{\|i - BMU\|^2}{2\sigma^2(t)}\right)$ , and  $L(t) = L_0 \exp\left(-\frac{t}{\eta}\right)$ .  $V(t)$  represents the input vector,  $\|i - BMU\|$  stands for the distance of the neuron  $i$  to the BMU, and the initial learning rate is  $L_0$ . To better understand how training in SOM works here is a brief summary of steps involved:

1. The weight for each node's are initialized.
2. From the set of training data, a vector is chosen at random.
3. Every node is examined to find the weights closest to the input vector. The winning node is known as the Best Matching Unit (BMU).
4. Calculate the neighborhood of the BMU. The number of neighbors would decrease over time.
5. The winning weight is rewarded with becoming more like the sample vector. The neighbors also get closer to the sample vector. Lastly, weights of the nodes closer to BMU get altered based on their closeness and the nodes that are farther away from BMU learns lesser.
6. Repeat step 2 for the chosen number of iterations.

After a number of iterations, the learning process starts to converge, and then the output neurons can be considered as representative of all the input vectors preserving the topology. As can be imagined this typical implementation of SOM can be utilized to reduce the colors to a few representative ones. However, it's not usually as simple since many medical images contain large amounts of noise in terms of various tissue, cells, nuclei, or glands, which may not be relevant for the problem being solved. SOM does provide a huge advantage in terms of easy interpretation and understanding of large amounts of data, which in this case is the confounded medical images. In addition, the grid clustering and reduction of dimensionality makes it easy to observe similarities in data.

### **3.5.3 Evaluation of Self Organizing Map**

As discussed, Self-Organizing Map allows for a graphical representation of output data in form of clusters that are created by reducing the dimensionality of input data on a resulting map. However, we need to make sure that these models are indeed a representation of the input data passed. To do so we need to evaluate the quality of these models which are represented in form of maps.

This quality of the Self-Organizing Map is generally measured using topology preservation and quantization precision. The latter is usually calculated as the squared quantization error, namely, the average distance between input vectors and corresponding best-matching units. It considers map to be better trained on given data based on how small the quantization error is. The former is based on the topology error, which is defined by the number of inputs to which the best-matching unit and the next best-matching unit are not adjacent on the map grid.

Although, if the true category of clusters is known then the confusion matrix serves as the most appropriate and commonly used technique for cluster evaluation. The confusion matrix helps in providing a more detailed evaluation of correct and incorrect classifications for each category. It detects the closeness of the composition of obtained clusters to that of the actual partition structure. Since neurons are more than actual clusters, a set of sub-clusters can be obtained while using SOM. In that case, the purity of the cluster is relevant for the resultant cluster, where the neurons of sub-cluster always belong to the member of the summarization cluster. Hence, each cluster is

identified as the dominating class label or the majority vote of its Voronoi set, while instances of other class is considered as error.

## 3.6 Conclusion

The above chapter has explained the various steps (shown in Figure 3.1) involved in the classification of medical data gathered from different sources while explaining the reasoning behind each step taken during the process. The list of sections starts by elaborating on the process involved in gathering knowledge from various sources and conducting interviews (see section 3.1) with experts in the medical field.

Furthermore, the details of various images including type, quality, and description are explained in data acquisition (see section 3.2). Pre-processing steps (see section 3.3) goes on to explains the steps involved in denoising the image by enhancing the relevant features that will be useful in later steps.

Section 3.4 explained the distinct methods, detailed algorithms, and provides illustrations of techniques used in representing these images. Finally, this chapter explained the classification and evaluation (see section 3.5) steps involved in analyzing the images and evaluating the results. There are multiple subsections providing the technical details and rationale for research progression. The implementation of the above discussed method and approach, along with different case studies to evaluate the performance is explained in chapter 4.

# Chapter 4

## Implementation and Case Studies

### 4.1 Introduction

This chapter provides a clear understanding of the classification of the medical image analysis system whose method was discussed in chapter 3. The sections in this chapter will provide details about the dataset used (see section 4.2). Include information on technical implementation of the system while showing the flow of data (see section 4.3). The section also explains the tools used, their system design, and specifications along with the information on the user interface. It also details the purpose and limitations of the tools used. In the later part of this chapter, we discuss the various case studies (see section 4.4), in which various different medical images were analyzed and their results are evaluated using the previously mentioned techniques. The system performance is discussed in the following section (see section 4.5) detailing the time taken for analysis and any related challenges that caused performance overhead. Finally, in the last chapter we discuss topic highlighting security and privacy concerns (see section 4.6) regarding the chosen implementation of this research and provide a brief conclusion in section 4.7.

### 4.2 Dataset Description

The details about the type of data and their sources have already been discussed in the above section 3.2.2. From that, it is already clear that the data includes a wide

variety of images that can be used in this implementation to evaluate the performance of the system for different case scenarios.

Even though there are various challenges in collecting medical images like unavailability, large size, privacy issues, etc. For the purpose of this research, we have collected data for 220 gastrointestinal images taken through the procedure of endoscopy. Out of these 220 images of endoscopy, there are 100 healthy gastrointestinal images (seen on top-left in figure 4.1) and 120 unhealthy gastrointestinal images (seen on top-right in figure 4.1) showing red spots that signifies bleeding region. These images are 256 x 256 pixels and as seen contain varying noise in form of food particles and body fluid.

The second set of images, taken from open source data provided by the Camelyon16

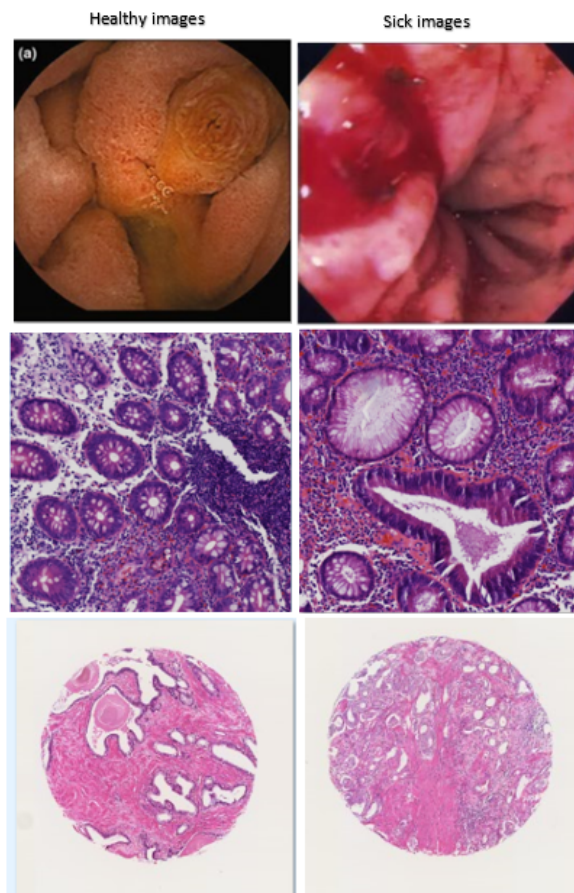


Figure 4.1: Healthy(on left) and Unhealthy(on right) images of endoscopy, lymph node tissue and prostate tissue in sequence from top to bottom.

challenge [92], consist of 1192 images cropped from the 400 whole slide images (WSIs) of the sentinel lymph node. The ratio of healthy to unhealthy images is 1:1 (i.e. there are 596 normal images and 596 images containing metastases). There are high resolution images with 1280 x 1280 pixel size. Lastly, the third set of images cropped from tissue samples taken 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. These include a total of 239 images of prostate tissue samples with 121 healthy samples and the remaining 118 images contain metastasis. The prostate tissue dataset was taken in two different settings and therefore include images with either 2000 x 2000 pixels or 5500 x 5500 pixels.

## **4.3 Technical Implementation and System Design**

In this section, the system architecture is discussed along with details of various tools used to provide end to end implementation. It also includes details of technical stack, programming languages, and the description of the user interface of systems that were utilized in this research.

### **4.3.1 System Architecture**

The system architecture of this implementation consists of multiple separate components between which the data is passed for processing, which is required in that step. A detailed diagram of system architecture is shown using figure 4.2. As it can be seen from the figure there are multiple components involved in the overall processing and starts with the dataset that is described earlier in section 4.2.

The details of other components such as Pre-processing script, Zen lite Application, CITU system and Matlab SOM are discussed below.

### **4.3.2 Technical Stack**

This section provides details and working of various script, tools, and technologies mentioned in system design architecture. It will cover the design, working, limitations, and functionalities of technologies used in this work.



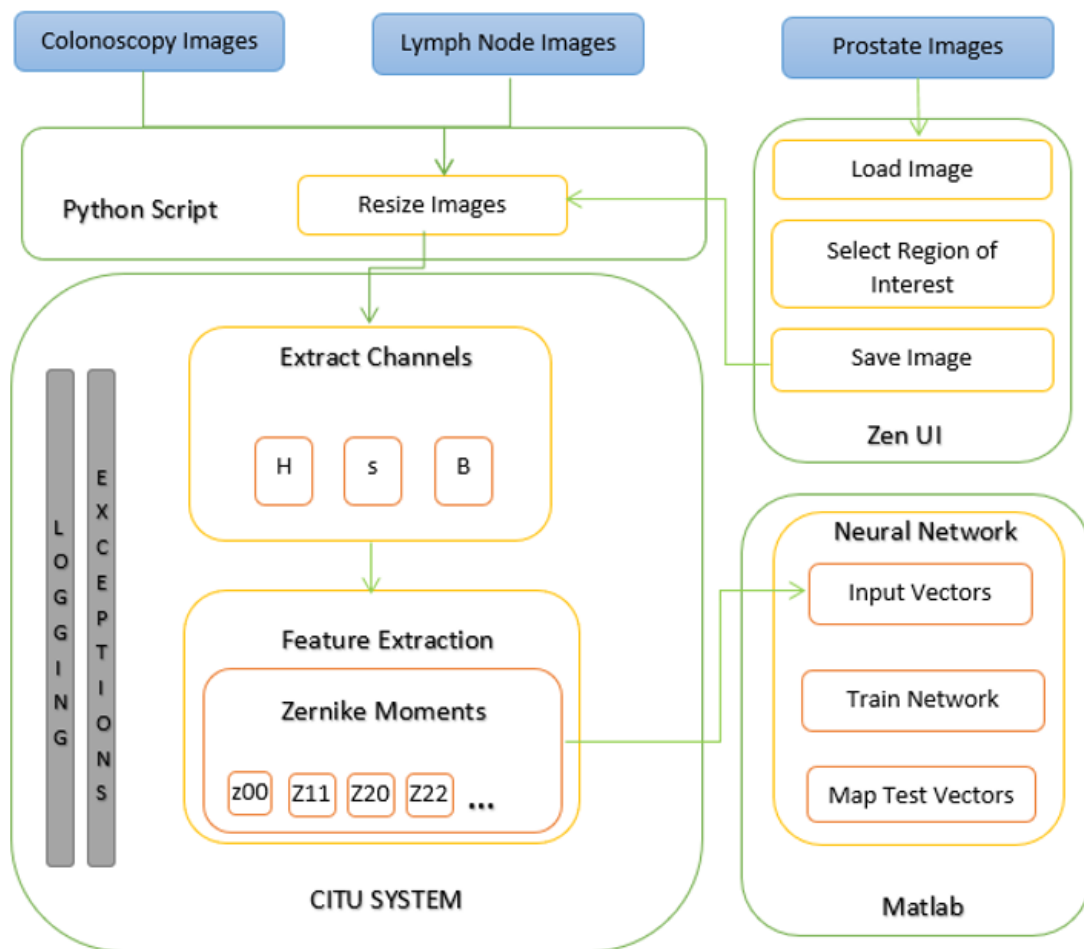


Figure 4.2: System Implementation Design for Medical Image Analysis.

### **Pre-processing Script**

This pre-processing script is used to resize the image into the required dimension of the same height and width. This step is very important in terms of how the data will be processed in later steps of image representation. This representation technique utilized the Zernike moments which defines a set of polynomials over the interior of the unit disc. The term ‘unit disc’ is important to note here. Due to this aspect of Zernike moment, the image is converted to the same length and width dimension for utilizing maximum area of analysis, therefore creating the requirement for this small script.

This script is written in Python language, which is a dynamic programming language that supports an extensive and powerful standard library with automated storage and realistic functionality. One such Python Imaging Library is called PIL, implemented by Fredrik Lundh and Contributors. This library contains an Image module that provides various functionalities including the option to resize the image. Using this option, the code iterates over the input folder location and parses over the list of all files available in that folder, while resizing them into the desired dimension of equal length and width. These files are saved in a new directory that is passed as a parameter to this code.

### **Zen Lite Application**

Zen, which stands for ZEISS Efficient Navigation, Lite application is the basic free version of the high-performance microscopy software ZEN. This system was created by Zeiss, which is one of the leading technology companies operating in the fields of optoelectronics and optics. This software solution works really well with the microscope images in CZI format.

The CZI format, which is again developed by ZEISS, is a well-known file format in microscope imaging as it stores imaging data along with all the relevant metadata in one compact file. These microscope images contain a huge amount of data and are usually large in size. So much so that normal image processing systems cannot even load these files let alone doing processing on them. In such situations, systems like Zen lite (shown in figure 4.3) can be really handy and act as a powerful tool for image analysis. Here is a screenshot of the ZEN user interface.

The ZEN user interface is quite simple but powerful. For this research, ZEN tool was used to analyze the large images (in size of gigabytes) of prostate tissue given by Dr. Aamir from Kings College of London. This tool provided the functionality

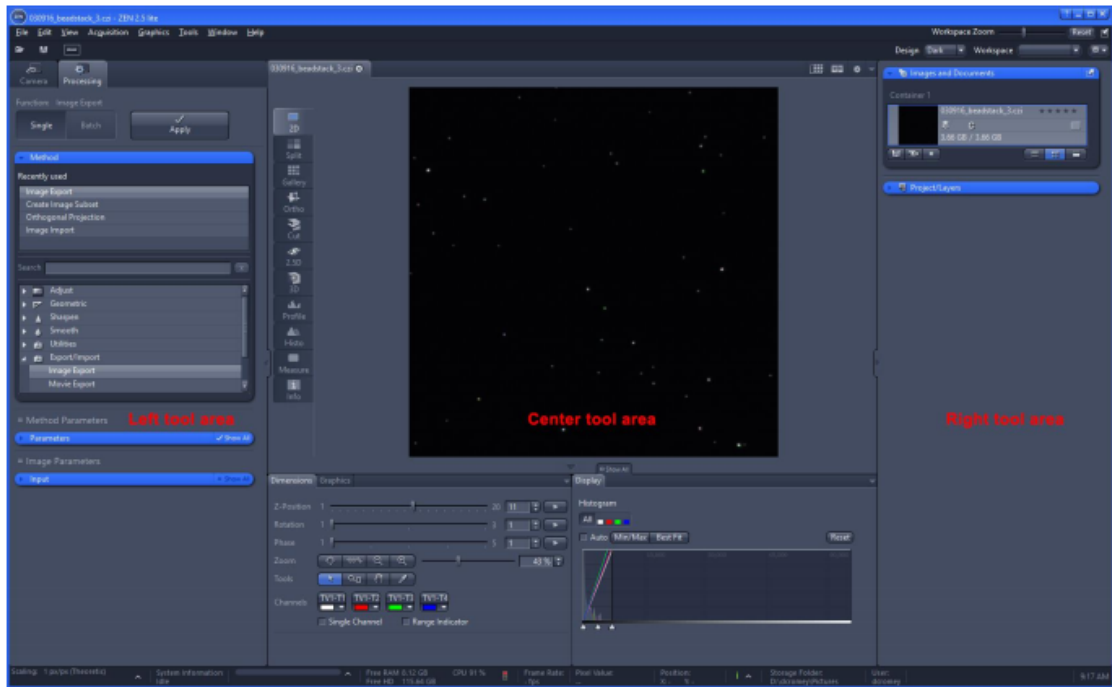


Figure 4.3: ZEISS ZEN lite software user interface.

of analyzing large size tissue images, by providing up to 20x zoom, and cropping the region of interest to save as healthy or diseased part of tissue to be included as a dataset image for later analysis.

### CITU system

CITU (Computerized Image and Text Understanding) system, which was developed at Trinity College Dublin under the supervision of Professor Khurshid Ahmad and obtained for the purpose of research, is a Windows based Graphical User Interface (GUI) application. The original intentions of developing this system was for text and image processing and perform cross-modal retrieval.

It has various other features that include image segmentation using techniques like Otsu thresholds, region growing techniques like the Watershed algorithm, and even self-organizing maps. There are also noise filtering techniques like Law's texture filter and Median noise filtering for providing a better quality of image. For the processing of text, there are functionalities available to compute frequency, TF-IDF, weirdness, and compound words are included. Moreover, to learn and establish associations between

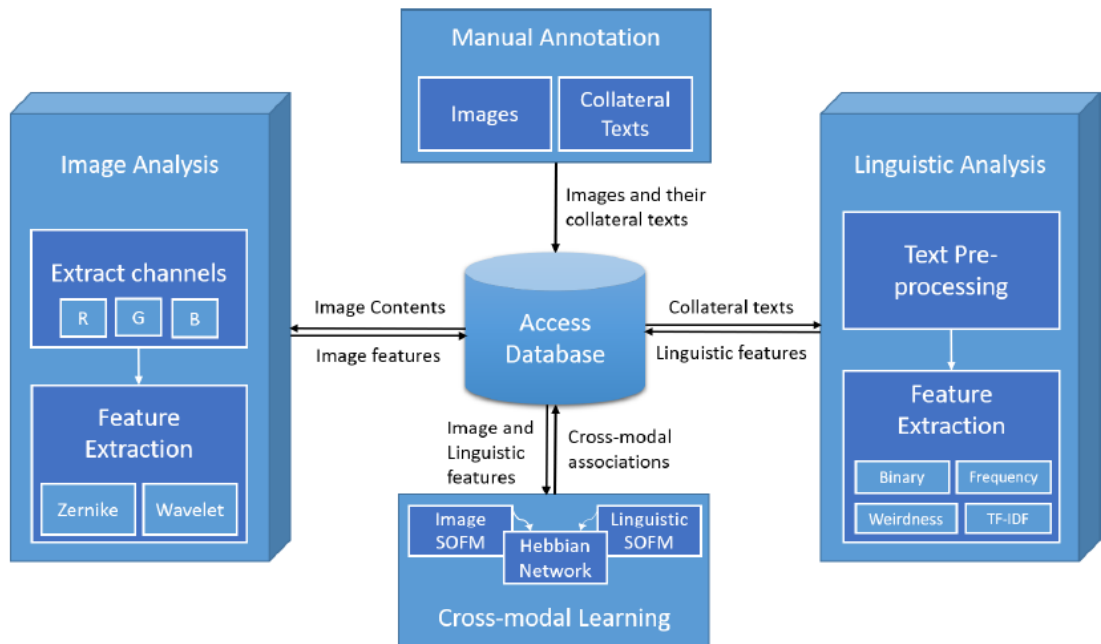


Figure 4.4: CITU system cross-modal design architecture

image features and linguistic features of annotated image data, a state-of-the-art cross-modal system is present as an option in Annotation tab of CITU. The architecture of CITU’s cross-modal functionality is presented in Figure 4.4.

CITU was written using C++ and requires installation and setup of Visual Studio and supporting packages for C++ project development. Apart from the linguistic analysis, CITU also supports image analysis under the feature extraction option of the Image tab. In Image analysis, CITU extracts the red, green, and blue channels from each image passed for analysis and performs a color transformation before extracting the Zernike feature of the passed color codes. In this case, the color transformation happens from RGB color scheme to HSV. Figure 4.5 shows the user interface for the CITU system.

The final results of extraction of RGB color, conversion to HSV and producing Zernike moments is saved into a text file on user picked location. CITU is also capable of processing multiple images at once by selecting the folder where all the images are stored and starting the above analysis by iterating over each file and storing results in the output file. CITU is a highly sophisticated system that supports many other

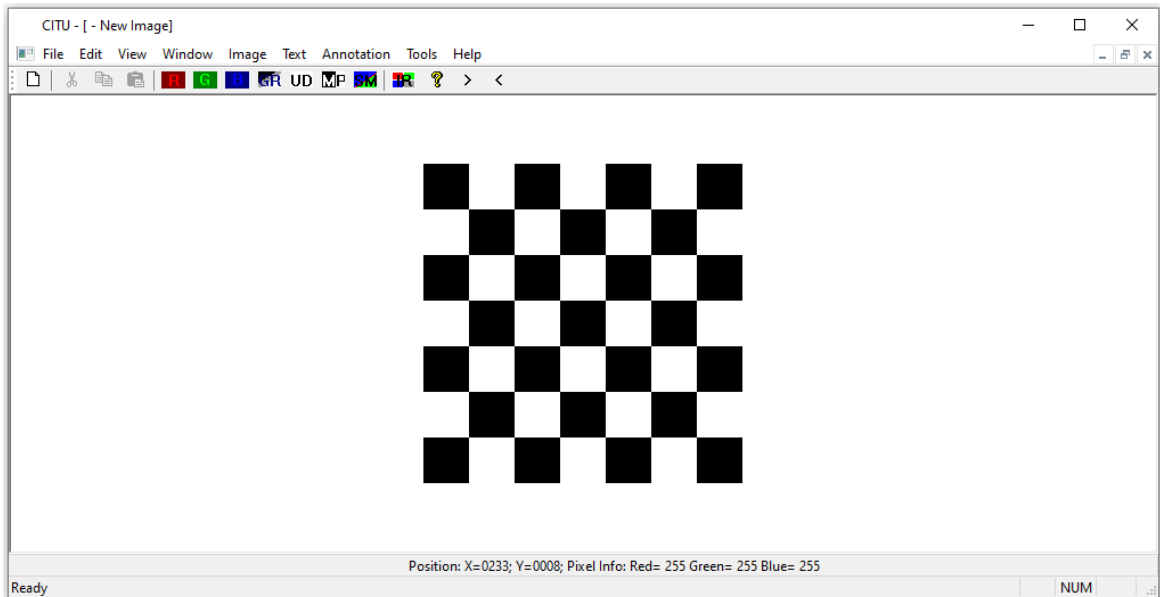


Figure 4.5: CITU system cross-modal design architecture.

features for text and image analysis which was not fully utilized here. However, in the future if the dataset is annotated or has linguistic features then CITU can perform an even more powerful analysis.

## MATLAB SOM

MATLAB, which stands for matrix laboratory, is a proprietary programming language that helps in the implementation of algorithms, plotting of functions and data, creation of user interfaces, matrix manipulation, and interfacing with programs written in other languages. This software was developed by MathWorks, which is an American privately held corporation. Its built-in graphics and ease of visualization makes it easy to gain insight from data. The version of software used for this research is MATLAB version 9.7 and its release name is R2019b.

MATLAB comes with various built-in apps amongst which is the Neural Network Clustering app (shown in figure 4.6) and can be seen under the apps tab, under Machine Learning. It can also be accessed using MATLAB command prompt, by typing the command `nctool` (to use the Neural Network GUI user can type `nstart`). The app version used in this research is Deep Learning Toolbox version 13.0. This neural net clustering app uses a self-organizing map to solves the challenges in the clustering

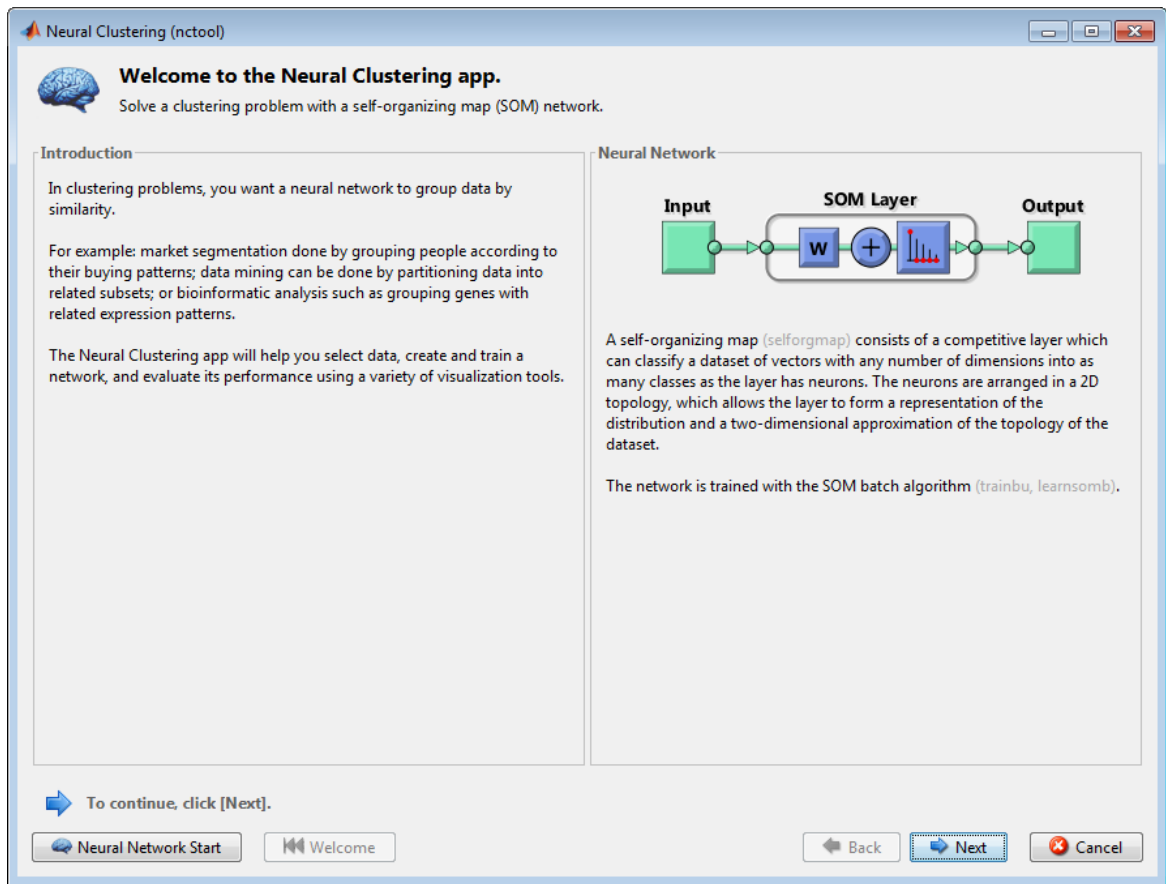


Figure 4.6: The GUI for Neural Network Clustering app in MATLAB.

problem. It allows user to choose their data, or even use one of their example datasets, and train the SOM network after configuring settings. Afterward one can even analyze the results using various visualization tools available, and if results are not satisfactory then even retrain the complete network with either updated settings or provide a larger dataset.

To start the analysis, we need to define the clustering problem. For this we can simply arrange the  $Q$  input vectors that needs to be clustered in a form of a matrix. For instance, if you have 10 two-element vectors then the input vector could be something like this.

$$\text{inputs} = [7 \ 0 \ 6 \ 2 \ 6 \ 5 \ 6 \ 1 \ 0 \ 1; 6 \ 2 \ 5 \ 0 \ 7 \ 5 \ 5 \ 1 \ 2 \ 2]$$

Matlab offers two types of algorithms for performing SOM, namely Sequential incremental SOM (the traditionally used method), and batch SOM. In the case of Sequential

incremental SOM, every input data vector is associated with the map unit, and using incremental training, with a learning rate factor, the map units converge towards the input data. While in the case of batch method, map units move closer to the data using Voronoi tessellation and, for this case, there is no learning rate that needs to be specified. By trying both the algorithm in [93] William et al. clearly indicated that the batch algorithm is faster by an order of magnitude. Moreover, it is also seen that the batch algorithm always results in the same solution with a specified set of parameters. But it is not the case with a sequential incremental SOM algorithm, as it tends to behave like a batch algorithm solution when the number of iterations is increased. However, repeated sequential algorithms do not result in the same solution every time. Because of this speed and reliability, it makes sense to use a batch algorithm for all the SOM analysis. Traditionally, the neural map, or network, has either a hexagonal, i.e. a honeycomb like structure, or a rectangular topology, i.e. trellis grid structure (seen in figure 4.7).

As can be seen in Figure 4.7, these topologies have different properties. In the case

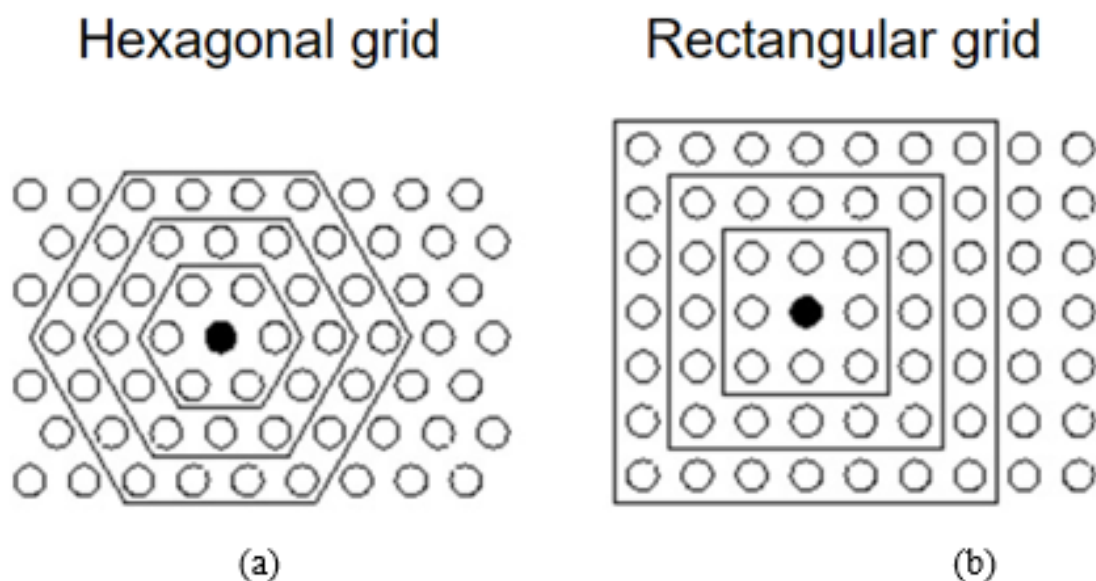


Figure 4.7: Represents the Topology structure of the Self-Organizing map grid, which traditionally is hexagonal or rectangular.

of hexagonal topology, Figure 4.7(a), there are six neighboring neurons for all the internal neurons. While for the rectangular topology, Figure 4.7(b), there are only four

neighboring neurons for all the internal neurons. However, the hexagonal topology is usually preferred because of its better visual effect [94] and its greater variance in neighborhood size. Another important factor is of network size while initializing Matlab SOM. The decision of finding the number of neurons to be used in the SOM must be made by the researcher and is usually the function of both size of the dataset and the purpose of application for which SOM is being used. But a basic thumb rule could be to use fewer neurons for applications requiring clustering of the dataset while using a larger number of neurons for visualization of high-dimensional datasets. To choose the number of iterations is also a required criterion for running Matlab SOM. This usually depends on convergence where a good map is one that is well converged and is representative of input data. It is usually done by running multiple tests on data to the find best results. However, it is important to point out that the neighborhood function, which in this case is the link distance function, strongly affect the number of iterations. So, fewer iterations might be required if the neighboring function is strong and the SOM becomes stable faster due to smoothing.

## 4.4 Case Studies and Discussion

The below section discusses about the analysis done on different medical images, namely endoscopy images, prostate tissue images and lymph node images. The detailed analysis is provided along with information on decisions made while choosing parameters and the impact of these decisions on overall evaluation.

### 4.4.1 Evaluation of Endoscopy Images

Analysis and evaluation of endoscopy images, discussed in section 4.2, collected from medical journals are provided in this section. Initially, the research started with finding a statistical relationship between the Zernike moments used to represent healthy and bleeding images. To do so Z-score of each moment for both categories(i.e. healthy and bleeding) of images was utilized. A Z-score is a numerical measurement that can be described as the input relationship to the mean of a group of values. It is a measurement in terms of standard deviations from the mean. Figure 4.8 below shows the comparison of these z-score values for healthy and bleeding endoscopy images in form



of histogram charts for different Zernike moments.

Observing these moments closely reveals that most of this data have a normal dis-

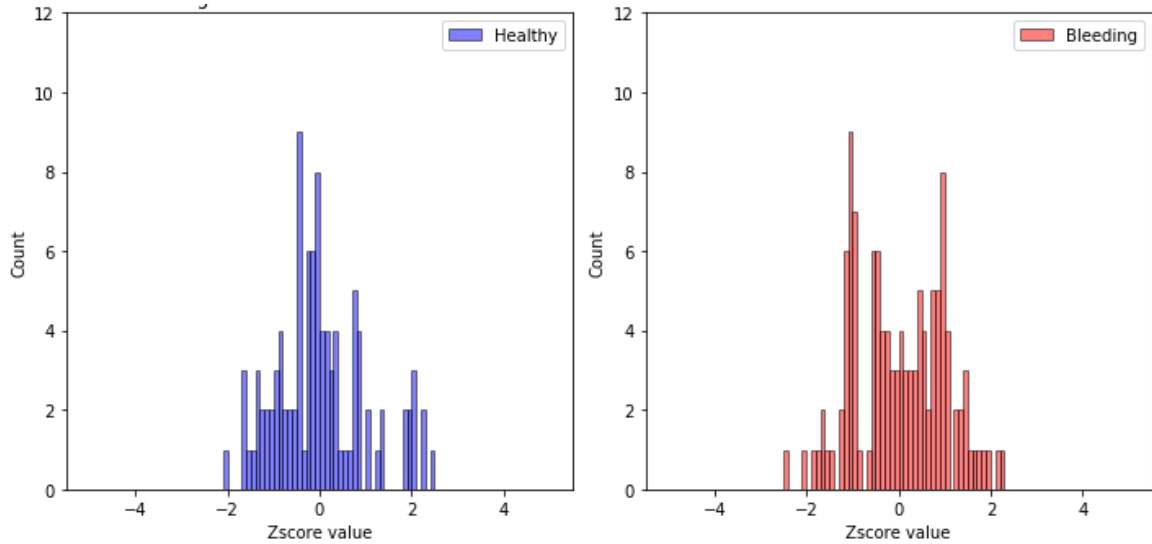


Figure 4.8: Graph showing z-score value values for z00 moment for healthy and bleeding endoscopy images.

tribution with values mostly lying up to 3 to 4 standard deviation from the mean. However, the curve is similar for each of the Zernike moment, suggesting that the data from mean shows the similar distribution. Seeing as the mean are also close for these moments of healthy and sick images it becomes really challenging to separate these data just based on the statistical analysis. Therefore, we resolve to more sophisticated methods of machine learning, which in this case is SOM.

With endoscopy image data, we divided the dataset into the ratio of 80-20 for training and test purpose. It is to be noted that while training no parameters were passed suggesting the category of data, only raw Zernike moments were utilized. After training the SOM with 80 healthy and 96 bleeding images over 1500 iteration the data is well distributed, except a higher concentration in the middle right, and shows the number of data points that are associated with each neuron (seen in Figure 4.9).

Another useful figure provided by Matlab SOM to analyze this training map is by using SOM Neighbor Distance figure (seen in figure 4.10). This figure allows us to understand the distances between neighboring neurons.

Color coding used in below figure is as follows:

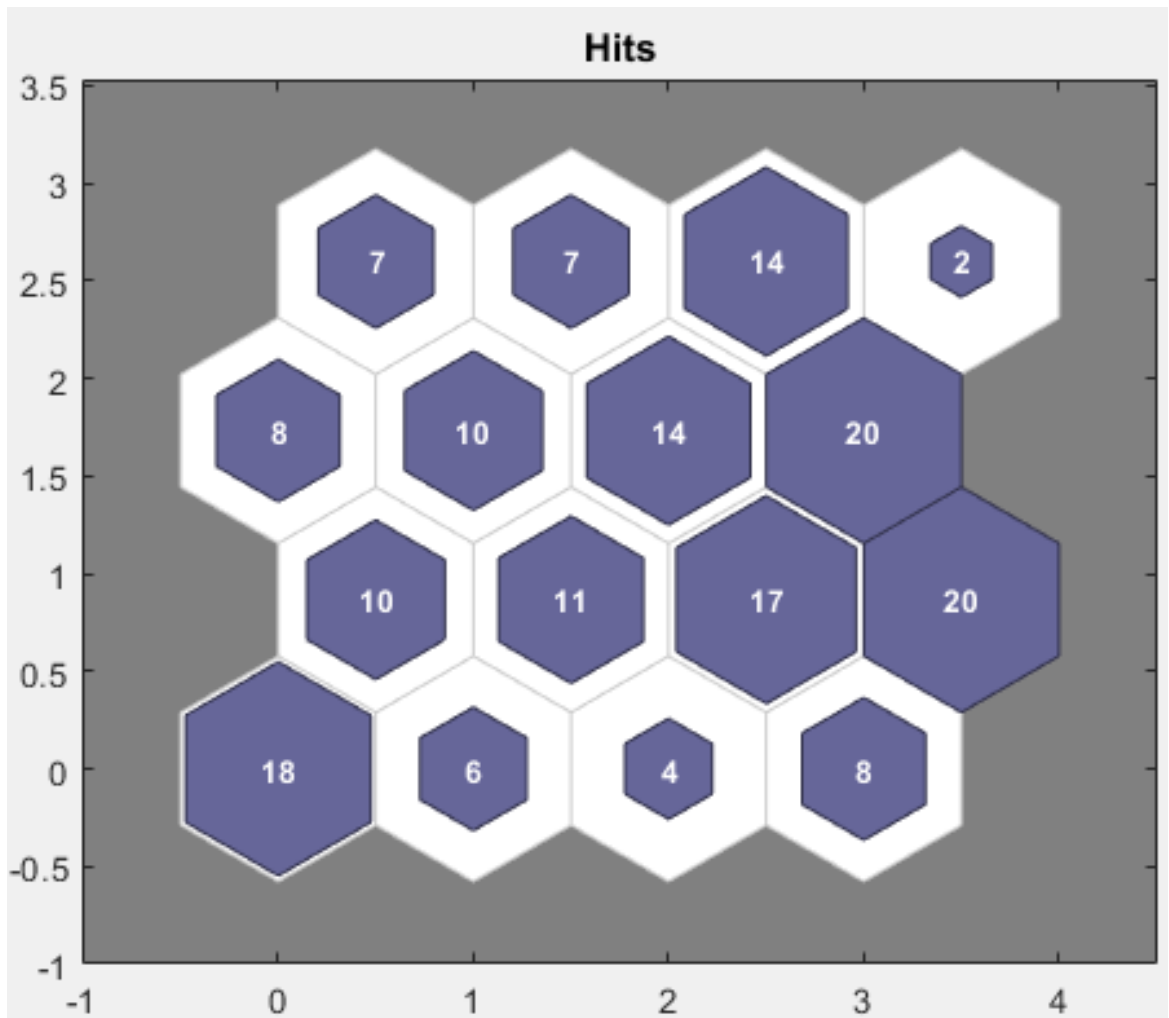


Figure 4.9: A 4x4 SOM sample hit map for endoscopy image dataset.

- Neurons are represented using the blue hexagons.
- Neighboring neurons are connected using the red lines between them.
- The distances between neurons is indicated by the colors in the regions containing the red lines.
- Larger distances are seen with a darker color.
- Smaller distance is shown using a lighter color representation.

Based on the above color-coding description we can analyze the Figure 4.10 showing a group of light segments appear in the middle-right region, bounded by some darker segments. This sort of grouping suggests that the map has clustered the given input data into groups. The similar grouping can be seen in Figure 4.9. The lower-left region in that figure also contains a group of tightly clustered data points. Their corresponding weights are also close in that region, as can be seen from the lighter color between neurons in neighbor weight distance figure. The region in the lower-right segment of the neighbor weight distance figure are darker than the ones on the middle left. This change in color also indicates that these regions are further apart.

Lastly, Matlab provides SOM weight planes to visualize the weight of input vectors using the weight plane figures. Figure 4.11 shows the weight planes for each of the input vector element, which in this case is 9 Zernike moments. These show the visualization of the weights that connect each input to the respective neuron. The value of weights is relative to the lighter and darker color of hexagon, where lighter suggests larger weights and darker color means smaller weights. They also show correlations between input variables that are very similar by presenting a very similar connection pattern. In this case, input 2 and input 5 show such correlation based on a similar color pattern. However, the rest of the inputs show a difference in weights and provide more knowledge.

Once the training is finished, we can test the remaining 20% data (i.e. 44 images) for finding the cluster positions based on the trained network. This test data is also passed as input vector and SOM returns an output vector which gives the position of the image in the cluster. The complete process to train and test this data took 5

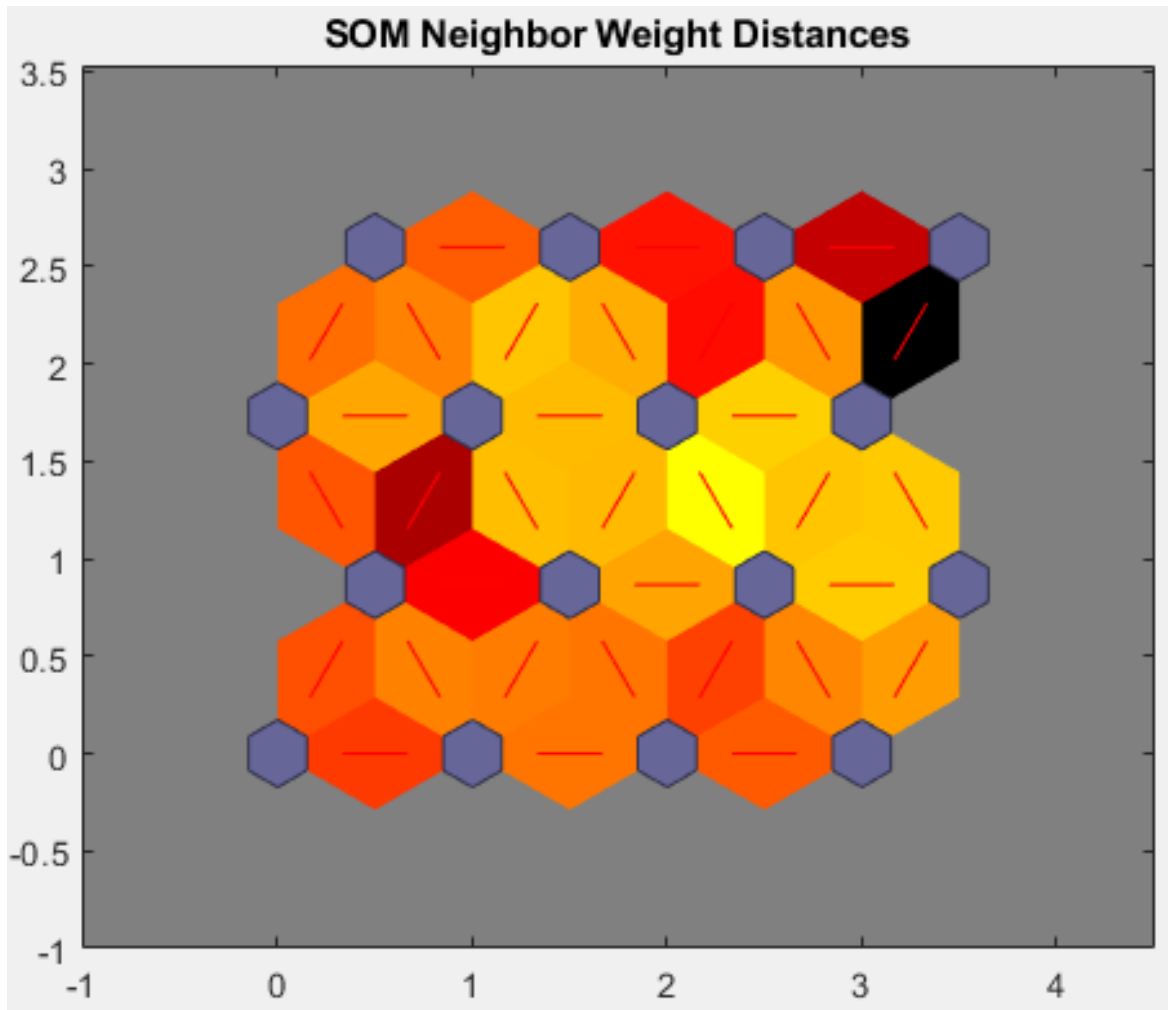


Figure 4.10: A 4x4 SOM neighbor weight distance graph for endoscopy image dataset.

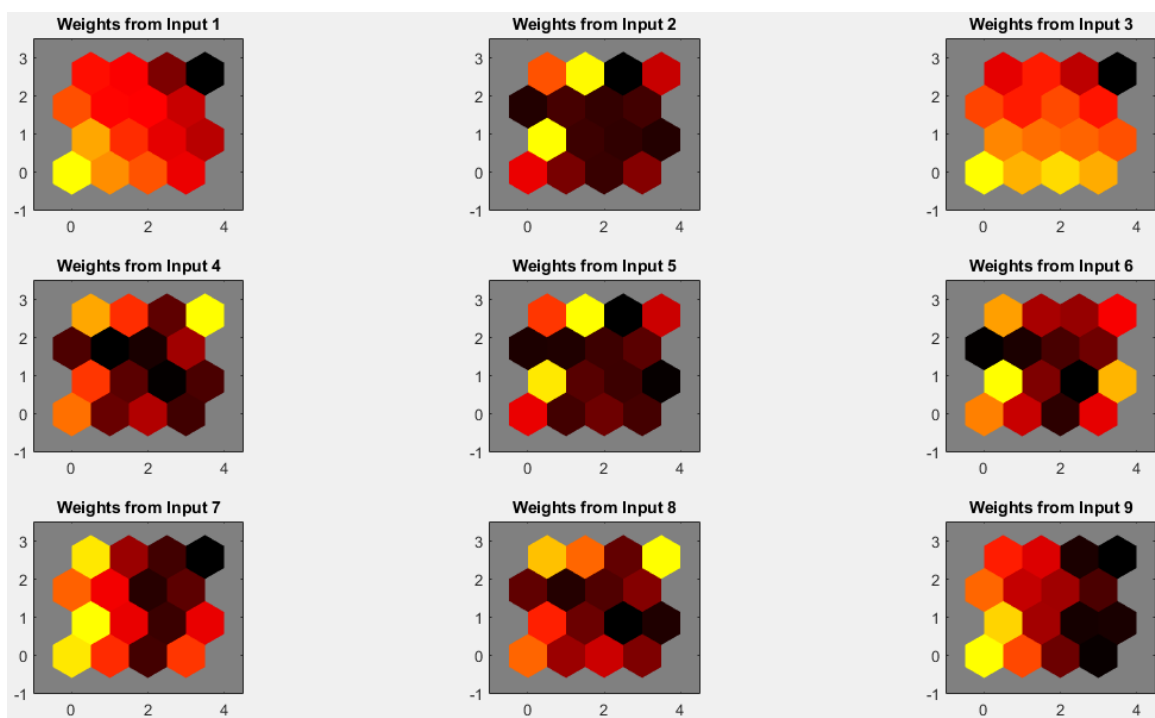


Figure 4.11: SOM weight planes for Endoscopy image data.

seconds using the mentioned approach.

Even though the data passed to SOM is unlabeled (as it uses unsupervised learning), we can utilize the evaluation technique, mentioned in section 3.5.3, to use the health condition labels available with images to assign these clusters as healthy or sick. Following this rule, the results of SOM network can be evaluated using a confusion matrix. In the case of endoscopy images, the resultant confusion matrix (as seen in Table 4.1) has a TP value of 17, FP is 1, FN is 7, and TN as 19. These terms, which are whole numbers, can be defined as:

- True positives (TP): These define the number which we predicted yes (they show the disease), and they do show the disease.
- True negatives (TN): Number of images predicted no, and they don't show the disease.
- False positives (FP): Number of images predicted yes, but they don't actually

have the disease.

- False negatives (FN): Number of images predicted no, but they actually do have the disease.

Table 4.1: Confusion matrix for cluster formed by endoscopy image dataset.

	True Condition			
	Images	Sick	Healthy	Total
Predicted	Sick	17	1	18
	Healthy	7	19	26
	Total	24	20	44

Using the confusion matrix, it is possible to calculate various other evaluation factors, one of which is accuracy. Accuracy can be defined as  $(TP+TN)/\text{total number of images}$ . The accuracy of endoscopy images using the above approach is 81%. Finally, Matlab can be used to visualize SOM sample hits on the test data. This visualization is on the same map that is used to train the network.

From the sample hit map shown in figure 4.12, we can clearly see the separation of two clusters, which also suggests a large distance between these two clusters of healthy and sick images. These clusters can be represented as holding healthy or sick images with few incorrectly assigned images that leads to false cases.

#### 4.4.2 Evaluation of Lymph Node Images

The analysis and evaluation techniques followed for lymph node images of breast cancer patients are discussed in this section and shows similar approach as taken in above section 4.4.1. It starts with representing images using Zernike moments after applying HSV color coding. For this analysis, the moments are taken up to order 5, which gives a total of 12 moments, for getting better results. However, it causes an increase in time taken for analysis. In this dataset, the time taken for complete evaluation is 12 second which is a 2x increase from the previous dataset. But there is also a large difference in

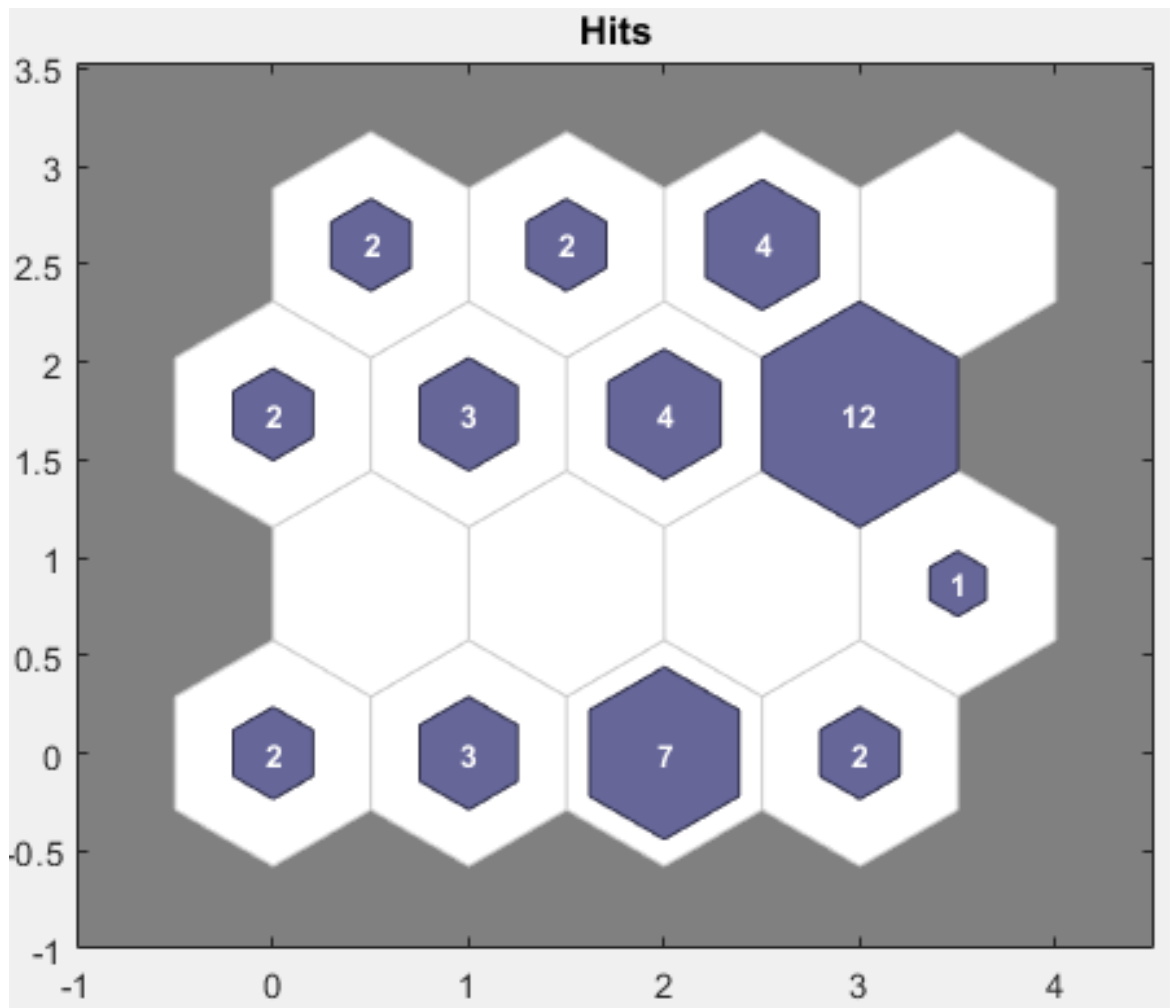


Figure 4.12: 4x4 test data SOM sample hit map for endoscopy image dataset.

the count of images used for analysis compared to the endoscopy dataset. Figure 4.13 shows the (a) SOM sample hit and (b) SOM neighbor weight distance visualization for the lymph node images.

In this case also we can observe a similar distribution of regions (one on left and

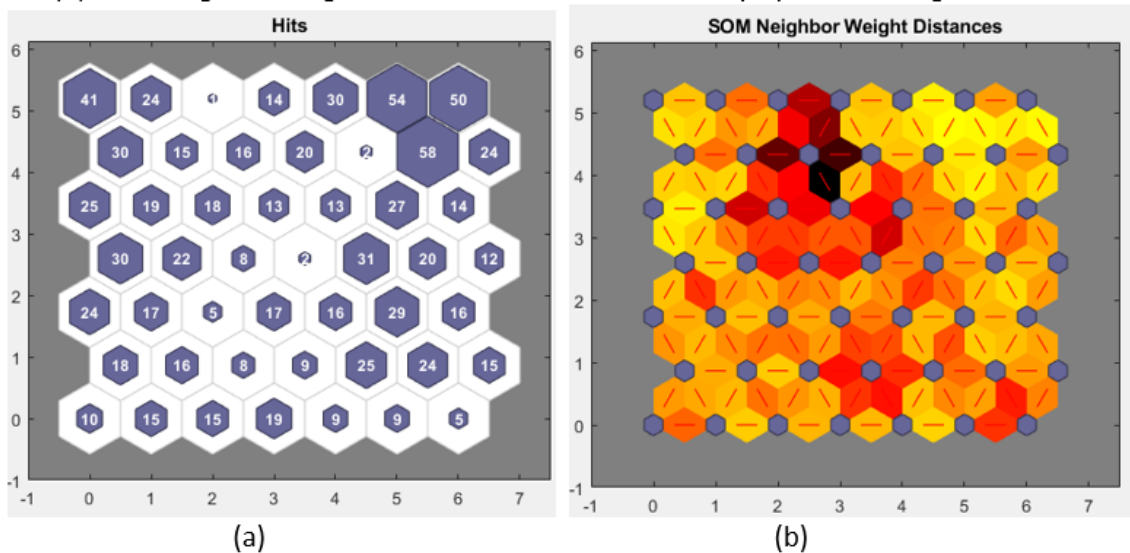


Figure 4.13: Represents a 7x7 (a) SOM sample hit and (b) SOM neighbor weight distance graph for the Lymph node image dataset.

the other on right) based on distance between the neighboring neurons (seen in figure 4.13b) and the color difference between them. Also, the overall spread in each neuron suggests a good distribution of data. SOM weight planes are shown using figure 4.14a below. Like before this distribution of weight planes have few similarities. However, the overall correlation is not very high, and each element of the vector can provide new knowledge. In this case study, the training dataset of 1192 images is divided into 80% training and 20% test data. The training is done using 1500 iterations and a learning rate of 0.9. The lymph node dataset is larger than the endoscopy one and shows a better convergence with a larger map size of 7x7. After training the test dataset shows the following results, using a confusion matrix, based on the cluster formed.

For the case of lymph node images, the resultant confusion matrix (as seen in Table 4.2) has TP value of 97, FP is 34, FN is 22, and TN as 85. This leads to an overall accuracy of 76% for lymph node images. Finally, figure 4.14b, shows the SOM sample



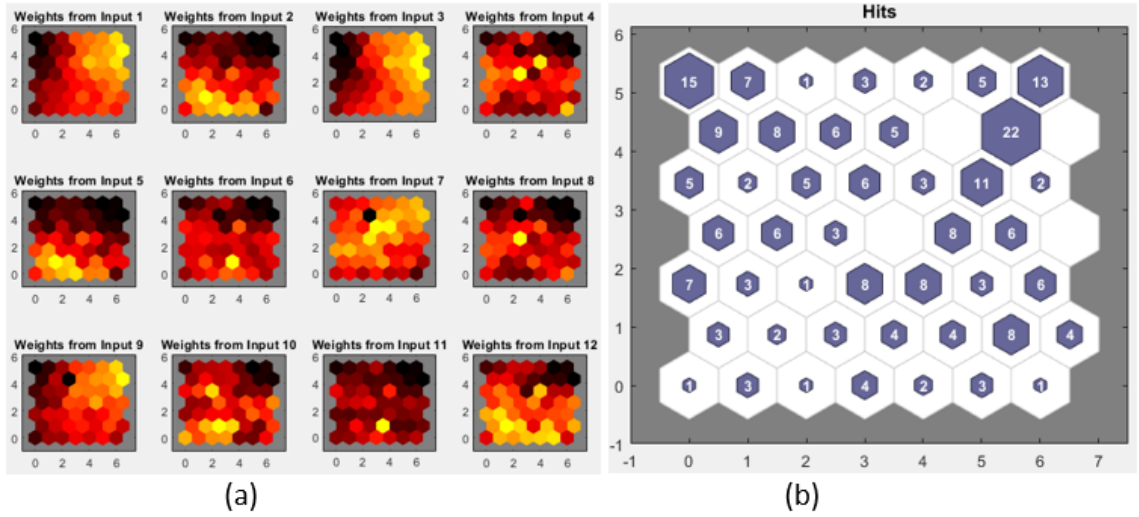


Figure 4.14: (a) SOM weight planes for 12 inputs of lymph node image train data. (b) SOM sample hit for lymph node image test data

Table 4.2: Confusion matrix for cluster formed by lymph node image dataset.

	True Condition			Total
	Images	Sick	Healthy	
Predicted	Sick	97	34	131
	Healthy	22	85	107
	Total	119	119	238

hit for lymph node test image dataset. Similar to the training map the test SOM sample hit also has a separation between cluster on the left and right side suggesting a clear boundary between healthy and sick images.

### 4.4.3 Evaluation of Prostate Tissue Images

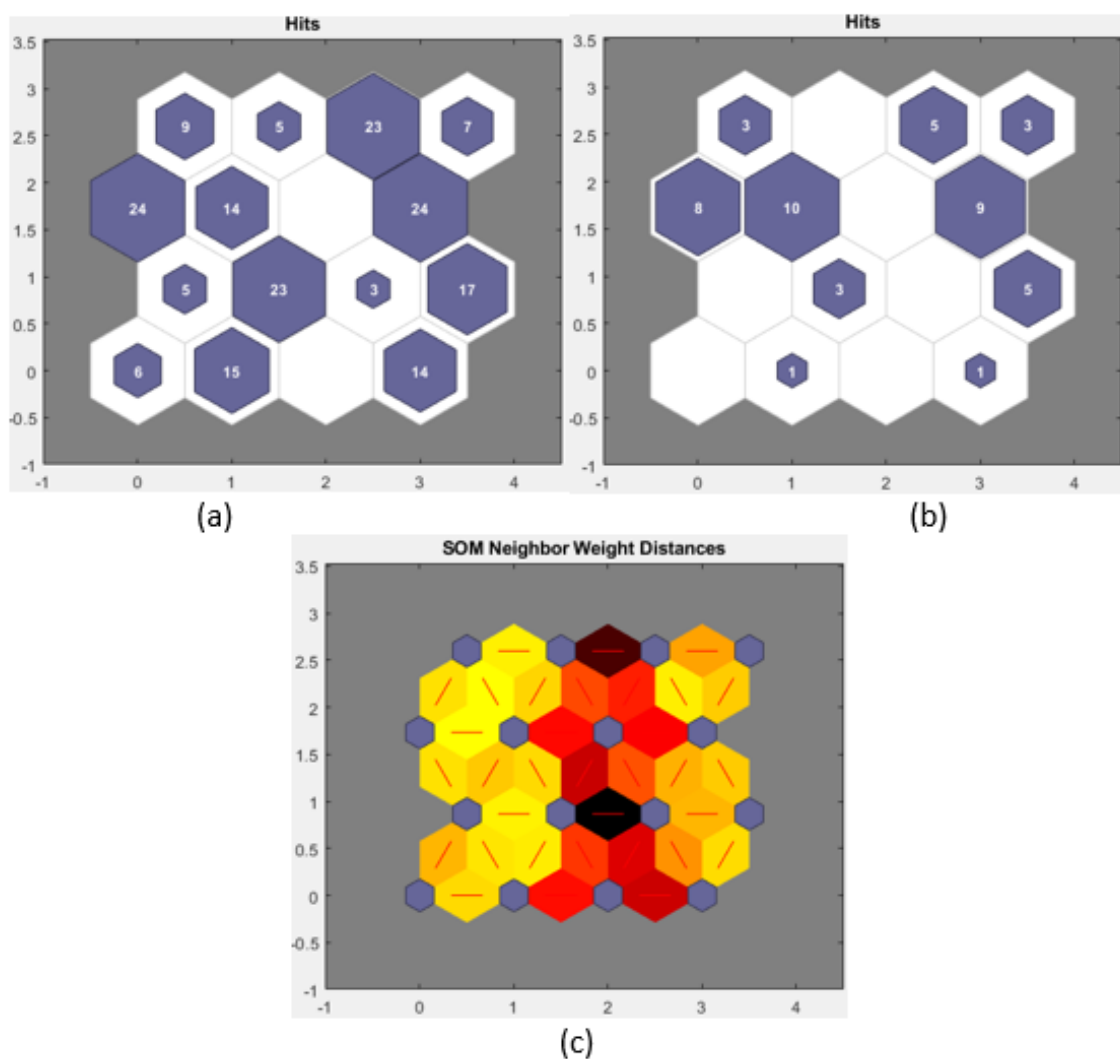


Figure 4.15: Represents a 4x4 (a) Training data SOM sample hit, (b) Test data SOM sample hit, and (c) SOM neighbor weight distance graph for Prostate tissue image dataset.

With prostate tissue images also, the dataset was divided between training and

testing in the ratio of 80-20, where 80% of data was used for training. The training procedure is similar to the one discussed in the above sections with 1500 iterations and a learning rate of 0.9. However, with prostate tissue images, there are two SOM map results that are presented to show the changes that happen when changing the map size. After representing the images using Zernike moment of order 4 the images are first trained using a 4x4 dimension SOM.

In Figure 4.15, the maps shown has a visible grouping that indicates that the network has clustered the data into two groups. The group on left side contains mostly healthy images, while the group on right represents a cluster of sick images. These groups are tightly coupled as visible from the neighbor weight distance graph. However, the boundary between them shows darker shade suggesting a larger distance between the neurons. The confusion matrix for this 4x4 dimension map is given in Table 4.3. Using the values given the accuracy of 4x4 dimensional map is 81%.

Table 4.3: Confusion matrix for cluster formed by prostate tissue image dataset.

	<b>True Condition</b>			
	<b>Images</b>	<b>Sick</b>	<b>Healthy</b>	<b>Total</b>
<b>Predicted</b>	<b>Sick</b>	16	2	18
	<b>Healthy</b>	7	23	30
	<b>Total</b>	23	25	48

Next, considering the same dataset of prostate tissue images and the above discussed process we train another map with a 5x5 dimension. Here the resulting maps are shown in Figure 4.16. As expected, the resulting maps are quite similar to the results seen in Figure 4.15 of the 4x4 dimension map. However, the boundary between the two groupings, the left and right grouping, is more significant and clusters now have a wider distance. Due to this increase in map size, there are more neurons and the data can spread more widely resulting in a smaller sized clusters. For this case, the grouping on left and right has completely separated the healthy and sick images with no overlapping images.

The weight plane graph below (seen in Figure 4.17) also shows a good relationship

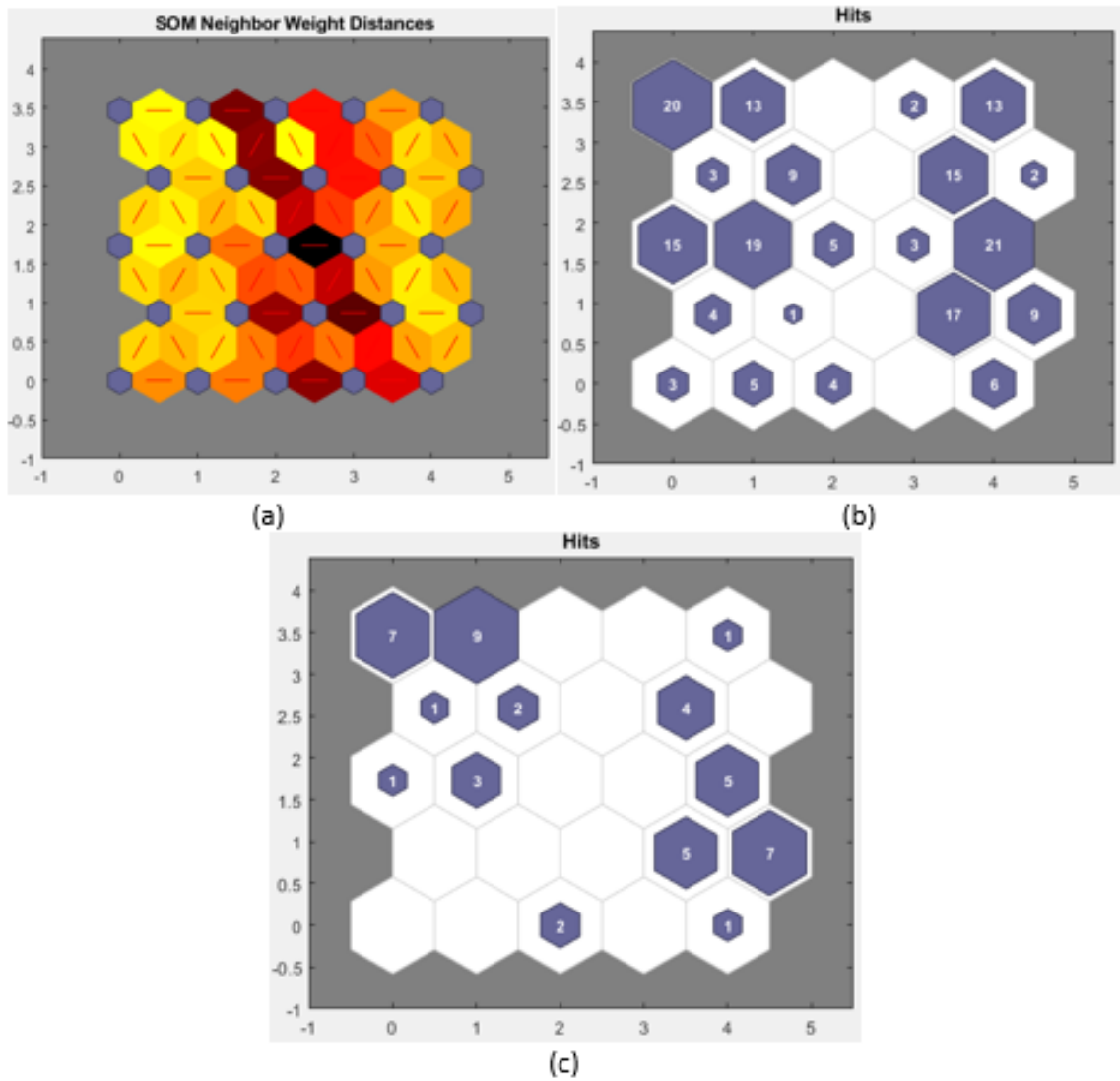


Figure 4.16: Represents a 5x5 (a) Training data SOM sample hit, (b) Test data SOM sample hit, and (c) SOM neighbor weight distance graph for Prostate tissue image dataset.

between vectors on different side of the graph and suggest a clear separation with others. Increasing the SOM map dimension from 5x5 to a 7x7 will separate the clusters even more by finding differences between neurons. Although, due to less data, lower differences, and large map size most of the neurons will be empty and the distribution

would not be even anymore. Hence, increasing the cluster size to 7x7 will not be recommended in this case.

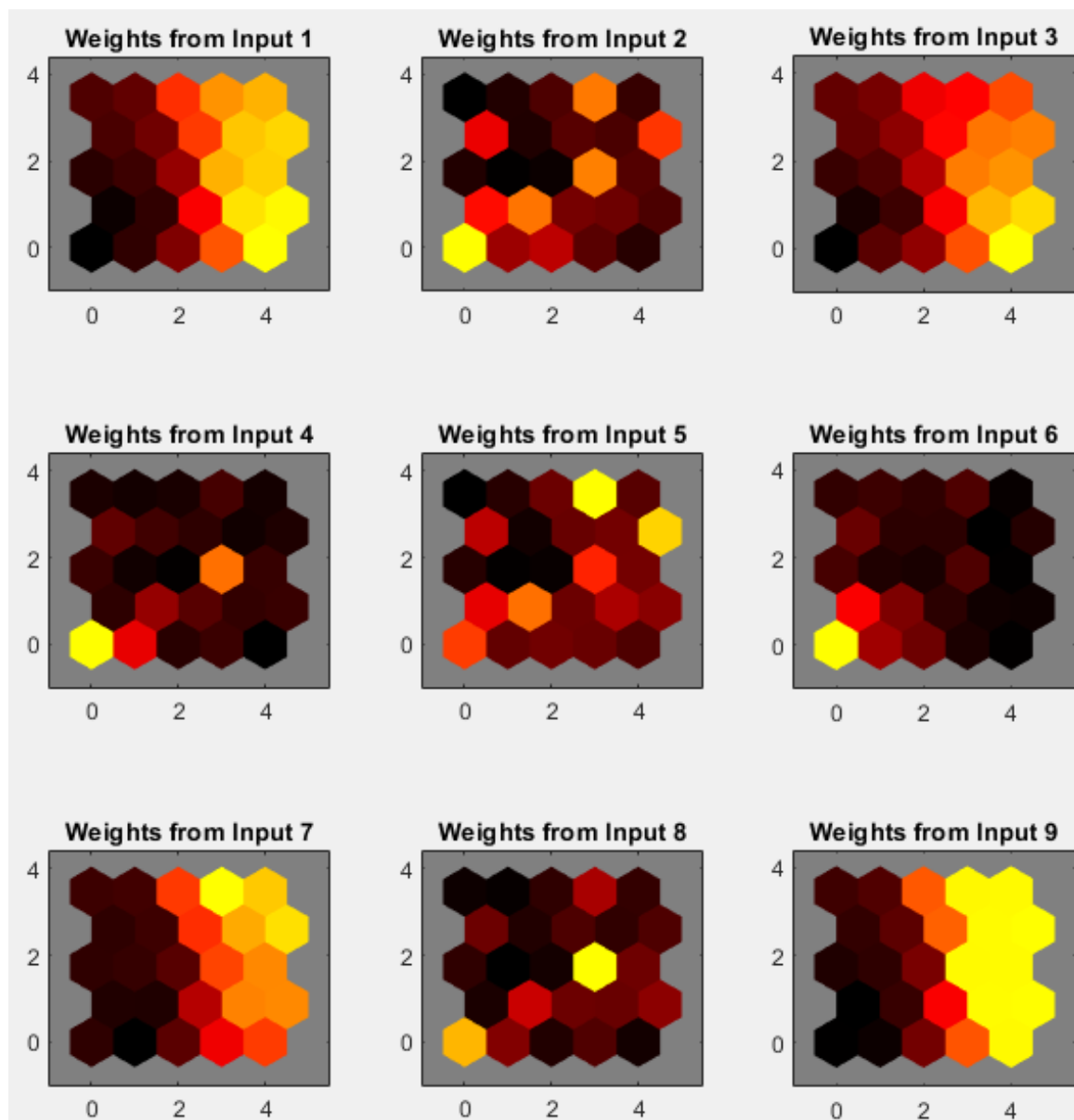


Figure 4.17: SOM weight planes graph for 9 inputs of prostate tissue image train data.

## 4.5 System Performance

In this research, the performance of the system varied based on the components defined in architecture. Since there were different datasets with multiple file sizes, image dimensions and the number of images, the system went through series of stress testing giving an idea about the performance of the system in different scenarios.

The system used for this experiment was a 64-bit Windows platform running over the Intel i7 processor with a 4.2 GHz processor. Based on these parameters the system was able to process the neural network clustering on the chosen datasets within 12 seconds for worst case scenario of 1192 images with 12 feature vectors for each image. The Zen Lite application was able to seamlessly load and process images with size up to 5GB for each image while handling UI. However, manual cropping takes time and has future scope for automation. Lastly, converting color code and calculating Zernike moment takes a fraction of a second for even large images. But when put together a set of thousands it could take up to a few hours to get numerical moments for all the images.

## 4.6 Security and Privacy Concerns

The analysis done in work deals with medical images of patients taken from various sources and utilizes free and open source tools (FOSS), as well as libraries, to process them. This leads to various concerns of security and privacy that we will discuss in this section.

### 4.6.1 Security Concerns

The security concerns revolving around this approach aims to prevent data and the code from being attacked or stolen. This could mostly be possible from any external input to system, which in this case is image data, or through some third-party software or library used in this system.

#### **Concerns with input image**

Any sort of input provides a gateway to the system and can be the weakest point in the

system if not handled well. Usually, these inputs are text based and most vulnerable to SQL injection type attacks. For this system, the input is an image file which can also be used to run a malicious script or download trojans if not handled carefully by the application.

In past there have been cases where simple PNG images contained malicious code that ran using technique known as RunPE, where the malicious code is executed in the context of another process. Similar other techniques have been used before and are difficult to detect even by the anti-virus systems.

A simple way to defeat such attacks would be re-encode an image, which is actually done in this approach. Taking the raw RGB values and then re-encoding those values to a new image in the same format can perform a validity check on the image and at the same time remove any junk code that might be attached to the image. However, there is still a scope of attacks such as the possibility of the image being a decompression bomb causing resource exhaustion DOS attacks. We can monitor the system resources in this case to prevent a crash or run this application with limited RAM (supported by CITU) so that application runs out of memory before crashing.

### **Concerns with third-party libraries**

Almost all applications used third-party libraries to boost their development process. While most of the critical vulnerabilities in third-party libraries are mentioned in the Common Vulnerabilities and Exposures (CVEs), the reality is most applications using these libraries do not handle these vulnerabilities well. Moreover, there could also be possibilities of unidentified vulnerabilities in these libraries.

To start handling these vulnerabilities the first step would be to list all the third-party software and libraries that are used while building the application. This information is already present in section 4.3 for this research. The third-party software list will allow us to find all the vulnerabilities available in these libraries along with their severity and probability of occurrence. Such information will prepare the developer in creating a plan to deal with known vulnerabilities. In this case, the third-party software used such as Matlab and Zen are more matured software that over the years have found known vulnerabilities and have released various security patches. However, the other components such as the python script and CITU system have been written without keeping security as a priority. To deal with their vulnerabilities there are code scanning tools,

like Bandit for python, that can scan each code file to find any sort of vulnerabilities that exists in these systems.

### 4.6.2 Privacy Concerns

As this system deals with various sort of medical image data, we will discuss in this section about the privacy concerns related to these medical images. As mentioned earlier, the images taken for endoscopy dataset are from published journals. Similarly, the images taken for the lymph node dataset are open source anonymized images. Since there is no personal information included on patients and the datasets are already published there is no privacy issue or ethics approval required.

The prostate 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. Institutional Review Board approved their use for research purposes. Only anonymized images from hematoxylin and eosin stained slides were used in this project. We did not use any ‘human tissue’, only images, so the ethical approval is not required. However, when using this system with non-anonymized medical images it is recommended to remove any personal information that might be related to the data. As this system only requires images and extracts data from it, there is no need to add any personal information of patients. Also, if there is any sharing of data over the network then a secure encryption technique should be utilized.

## 4.7 Summary

In the above chapter we discussed in detail about the dataset (see section 4.2) used for analysis along with the architecture of the system (see section 4.3) that analyzes this data. The details of tools and third-party libraries are also mentioned. To show the viability of this system with real world data and having the potential to be used in clinical practice, this research presents three case studies (see section 4.4) done on multiple medical image datasets with promising results. The system performance is also discussed in section 4.5 to address the scalability and limitation when it comes to testing the checking system threshold. Lastly, the security and privacy concerns (see



section 4.6) talks about the vulnerabilities and issues that could occur while using this system in the real clinical scenarios. It also mentions the solutions for such problems.

# Chapter 5

## Conclusion and Future Work

This chapter presents the conclusion of the undertaken work as part of this dissertation. It includes the key take away of this work and some weaknesses when using this approach. To fix some of these weaknesses the second subsection lists future works that can be implemented to improve or extend this research.

### 5.1 Conclusion

Medical image processing is a complex technique comprising of interdisciplinary fields like mathematics, computer science, medicine and physics. In this work, a medical image analysis algorithm is designed and implemented to categories the medical images into different clusters based on their characteristic visual features. This categorization of medical images is performed by training an unsupervised network of Kohonen's self-organizing maps that leverages the similarity of input vector to produce a low dimensional representation of input space, called map. To achieve this the input given to the system is required in form of numerical representation of these medical images. After going through various available techniques of image representation, this work takes advantage of Frits Zernike's approach of using sequence of polynomial called Zernike moments to capture the properties of an image. Since these polynomials are orthogonal to each other, the information captured from these images using Zernike moment is not redundant or has any overlapping information which may cause some sort of bias in training. The presented work can contribute to the field of medical

science by providing a technique capable of analyzing a large number of images in a very short span of time without using any label on the input image dataset. However, this system is not a replacement to human pathologists that analyze these images in research labs, rather a complimenting tool that can predetermine the data to analyze, categorize them into relevant sections, provide the second opinion or do a preliminary analysis.

There are details about system architecture, algorithms used, and information on

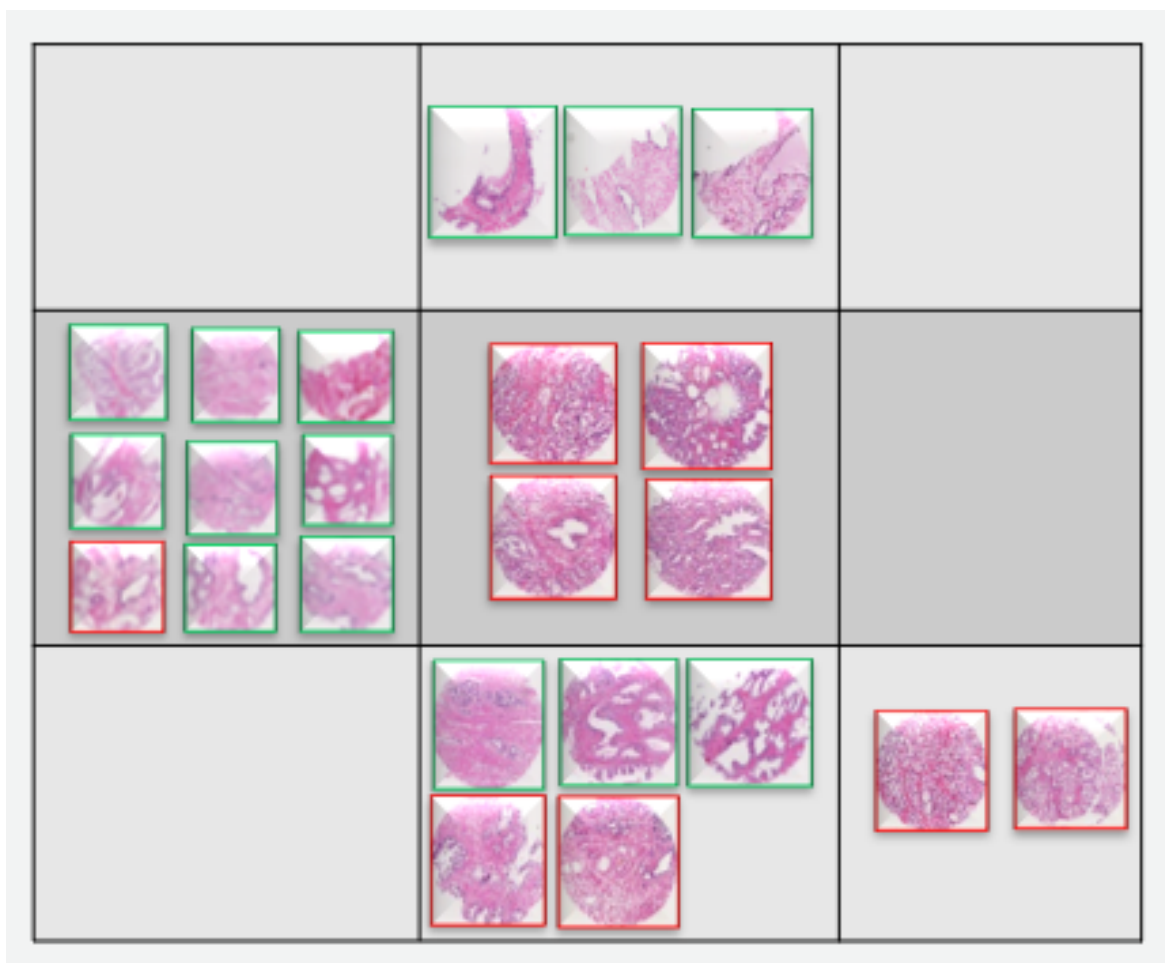


Figure 5.1: A 3x3 Map showing clustering of healthy (marked with green) and sick (marked with red) of prostate tissue samples.

various tools or languages used to implement this system. The advantages and short-coming are also discussed when using these approaches. Few case studies are presented

for running, testing, and evaluating the given approach with real world datasets. These studies showcase the potential such systems have in the future of medical image processing.

Using the mentioned implementation, the case studies are performed on 3 medical image dataset of endoscopy images, lymph node images and prostate tissue images. After the pre-processing and analysis, the endoscopy and prostate tissue images cluster into sick and healthy image with an accuracy of 81%. While the lymph node image clustering has 76% accuracy. However, the results may vary with map size and for a more optimal approach, a Hierarchical SOM (HSOM) or Growing SOM (GSOM) could be utilized. These maps also provide placement of images on a two-dimensional grid which allows a view of the complete database in one image, as seen in Figure 5.1. Finally, this work can be adapted to more than just the chosen medical image dataset as it's not based on any specific feature of tissue or sample structure.

## 5.2 Future Work

Analyzing the results and method of implementation there are few changes that could have improved the results while others could have handled the limitations observed in this work. These changes could be expanded to the way the data is handled and how the processing on this data could be made more efficient. Future work for this system can include the below changes by making an extension to the program that is more enhanced and can handle more complex data.

1. One of the biggest challenges with the medical images is the large size of data. Specially in case of whole slide images the data could reach into range of tens of gigabytes. When processing this data, the system usually requires high RAM and take enough time to process it. In this case, the CITU system and python script have shown poor performance when going through such large images. To improve this there could be few solutions like doing parallel processing with python script, utilizing the map reduce framework of big data (require lot of code changes), or consider cloud computing platform services. The third option of cloud computing platform service like docker seems more suitable due to its capacity to scale (by running more docker instance) based on requirement without need to write extra

pieces of code.

2. The second challenge is like the above as it is due to large file size image but deals with finding region of interest. Since the image file sizes are large, they need to be cropped into smaller segments that are easier to process and analyze. To achieve this, a manual process of finding the region of interest and cropping is required. Usually done with tools like Zen lite that are capable of loading such large files. However, a simple automation solution like a python code to split file into required segments would work here. Although the changes made must be done in a way that the code is able to handle such large file size.
3. The process of manually analyzing the SOM map size is inefficient. Even if the best map size is found, the time take to achieve this is significant. However, it does give a better understand on how the data reacts to different map sizes. To this more effectively there are various available techniques like Hierarchical self-organizing map (HSOM) or Growing self-organizing map (GSOM) which can automatically analyze the more effective map size for the chosen dataset.
4. Medical image data are not always perfect and can even have problems due to errors in image capturing process. In such cases it becomes challenging to capture the features of image based on just the visual features of the image. In such cases linguistic cues from experts can be added with images to provide extra set of features that can help in identifying the image more accurately. Such a system multi modal system can even be used in content-based image retrieval process.

The above suggested future work is based on limitations observed and are provided with valid solutions. However, in future there is a possibility of more challenges being discovered. Therefore, it's important to note that while implementing any solution there are few things that should be kept in mind like the system performance due to changes, the security and privacy concerns it might have and to mention any new limitations that the changes might introduce.

# Bibliography

- [1] “List of qualified biomarkers.” <https://www.fda.gov/drugs/cder-biomarker-qualification-program/list-qualified-biomarkers>. Accessed: 2020-08-28.
- [2] D. Stipanicev, “Introduction to digital image processing and analysis,” 1994.
- [3] M. Gamarra, E. Zurek, and H. San Juan Vergara, “Study of image analysis algorithms for segmentation, feature extraction and classification of cells,” *Journal of Information Systems Engineering Management*, vol. 2, 08 2017.
- [4] C. Schneider, W. Rasband, and K. Eliceiri, “Nih image to imagej: 25 years of image analysis,” *Nature Methods*, vol. 9, 07 2012.
- [5] M. R. Zare, W. C. Seng, and A. Mueen, “Automatic classification of medical x-ray images,” *Malaysian Journal of Computer Science*, vol. 26, 2013.
- [6] Y.-D. Zhang, Z. Dong, L. Aijun, S. Wang, G. Ji, Z. Zhang, and J. Yang, “Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine,” *Journal of Medical Imaging and Health Informatics*, vol. 5, 11 2015.
- [7] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, “Two systems for the detection of melanomas in dermoscopy images using texture and color features,” *IEEE Systems Journal*, vol. 8, no. 3, pp. 965–979, 2014.
- [8] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” pp. 844–848, 2014.

- [9] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. J. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *CoRR*, vol. abs/1602.03409, 2016.
- [10] E. Yu, A. Basavanthally, J. Xu, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, “Incorporating domain knowledge for tubule detection in breast histopathology using O’Callaghan neighborhoods,” in *Medical Imaging 2011: Computer-Aided Diagnosis* (R. M. S. M.D. and B. van Ginneken, eds.), vol. 7963, pp. 305 – 319, International Society for Optics and Photonics, SPIE, 2011.
- [11] A. Janowczyk, J. Xu, S. Chandran, and A. Madabhushi, “A weighted mean shift, normalized cuts initialized color gradient based geodesic active contour model: Applications to histopathology image segmentation,” *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 7623, 03 2010.
- [12] H. Shimada, O. Sertel, U. V. Catalyurek, and M. N. Gurcan, “Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1433–1436, 2009.
- [13] O. S. Al-Kadi, “Texture measures combination for improved meningioma classification of histopathological images,” *Pattern Recognition*, vol. 43, no. 6, pp. 2043 – 2053, 2010.
- [14] M. Kandemir, C. Gunduz-Demir, A. B. Tosun, and C. Sokmensuer, “Automatic segmentation of colon glands using object-graphs,” *Medical Image Analysis*, vol. 14, no. 1, pp. 1 – 12, 2010.
- [15] J. Chatterjee, M. Muthu Rama Krishnan, M. Pal, S. K. Bomminayuni, C. Chakraborty, R. R. Paul, and A. K. Ray, “Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis—an svm based approach,” *Computers in Biology and Medicine*, vol. 39, no. 12, pp. 1096 – 1104, 2009.
- [16] M. Gurcan, H. Kong, and K. Belkacem-Boussaid, “Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and

- cell splitting,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1661–1677, 2011.
- [17] M. N. Gurcan, M. M. Dondar, S. Badve, G. Bilgin, V. Raykar, R. Jain, and O. Sertel, “Computerized classification of intraductal breast lesions using histopathological images,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 7, pp. 1977–1984, 2011.
- [18] 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, pp. 127rv3–127rv3, 2012.
- [19] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, “Breast cancer histopathology image analysis: A review,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.
- [20] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, “Histopathological image analysis: A review,” *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.
- [21] A. Belsare, “Histopathological image analysis using image processing techniques: An overview,” *Signal Image Processing An International Journal*, vol. 3, no. 4, pp. 23–36, 2012.
- [22] N. Stathonikos, M. Veta, A. Huisman, and P. van Diest, “Going fully digital: Perspective of a Dutch academic pathology lab,” *Journal of Pathology Informatics*, vol. 4, no. 1, p. 15, 2013.
- [23] K. T. Ahmed, S. Ummesafi, and A. Iqbal, “Content based image retrieval using image features information fusion,” *Information Fusion*, vol. 51, pp. 76 – 99, 2019.
- [24] S. H. Shirazi, A. I. Umar, S. Naz, N. ul Amin Khan, M. I. Razzak, and B. Al-Haqbani, “Content-based image retrieval using texture color shape and region,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016.



- [25] S. R. Dubey, S. K. Singh, and R. Kumar Singh, “Local neighbourhood-based robust colour occurrence descriptor for colour image retrieval,” *IET Image Processing*, vol. 9, no. 7, pp. 578–586, 2015.
- [26] M. Verma and B. Raman, “Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval,” *Multimedia Tools and Applications*, pp. 1–25, 05 2017.
- [27] X. Duanmu, “Image retrieval using color moment invariant,” in *2010 Seventh International Conference on Information Technology: New Generations*, pp. 200–203, 2010.
- [28] V. H. Vu, Q. N. Huu, and H. N. T. Thu, “Content based image retrieval with bin of color histogram,” in *2012 International Conference on Audio, Language and Image Processing*, pp. 20–25, 2012.
- [29] 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, 01 2018.
- [30] D. Zhang and G. Lu, “Review of shape representation and description techniques,” *Pattern Recognition*, vol. 37, no. 1, pp. 1 – 19, 2004.
- [31] C. Di Ruberto, A. Loddo, and L. Putzu, “Histological image analysis by invariant descriptors,” in *Image Analysis and Processing - ICIAP 2017* (S. Battiato, G. Gallo, R. Schettini, and F. Stanco, eds.), (Cham), pp. 345–356, Springer International Publishing, 2017.
- [32] K. Wu, C. Garnier, J. 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*, pp. 5581–5584, 2010.
- [33] 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.

- [34] C. Zheng, A. Long, Y. Volkov, A. Davies, D. Kelleher, and K. Ahmad, “A cross-modal system for cell migration image annotation and retrieval,” pp. 1738 – 1743, 09 2007.
- [35] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” *PLOS ONE*, vol. 12, pp. 1–14, 06 2017.
- [36] H. Källén, J. Molin, A. Heyden, C. Lundström, and K. Åström, “Towards grading gleason score using generically trained deep convolutional neural networks,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 1163–1167, 2016.
- [37] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” pp. 1626–1630, 05 2014.
- [38] D. Karimi, G. Nir, S. Hor, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, “Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts,” *Medical Image Analysis*, vol. 50, pp. 167 – 180, 2018.
- [39] J. E. Tomaszewski, J. P. Monaco, M. D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, “High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models,” *Medical Image Analysis*, vol. 14, no. 4, pp. 617 – 629, 2010.
- [40] A. Madabhushi, S. Doyle, M. Feldman, and J. Tomaszewski, “A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1205–1218, 2012.
- [41] M. Gaed, L. Gorelick, O. Veksler, J. A. Gómez, M. Moussa, G. Bauman, A. Fenster, and A. D. Ward, “Prostate histopathology: Learning tissue component his-

- tograms for cancer detection and classification,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 10, pp. 1804–1818, 2013.
- [42] A. Sarkar, K. Nguyen, and A. K. Jain, “Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2254–2270, 2014.
- [43] E. Arvaniti, K. S. Fricker, M. Moret, N. J. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, and M. Claassen, “Automated gleason grading of prostate cancer tissue microarrays via deep learning,” *bioRxiv*, 2018.
- [44] J. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. Zehnbauer, K. Lister, and R. Parwaresch, “Breast carcinoma malignancy grading by bloom-richardson system vs proliferation index: Reproducibility of grade and advantages of proliferation index,” *Modern Pathology*, vol. 18, pp. 1067–1078, Aug. 2005.
- [45] E. A. Perez, V. J. Suman, N. E. Davidson, S. Martino, P. A. Kaufman, W. L. Lingle, P. J. Flynn, J. N. Ingle, D. Visscher, and R. B. Jenkins, “Her2 testing by local, central, and reference laboratories in specimens from the north central cancer treatment group n9831 intergroup adjuvant trial,” *Journal of Clinical Oncology*, vol. 24, no. 19, pp. 3032–3038, 2006. PMID: 16809727.
- [46] J. Kim and J. F. Courtney, “A survey of knowledge acquisition techniques and their relevance to managerial problem domains,” *Decision Support Systems*, vol. 4, no. 3, pp. 269 – 284, 1988.
- [47] A. J. Symes, M. Eilertsen, M. Millar, J. Nariculam, A. Freeman, M. Notara, M. R. Feneley, H. R. H. Patel, J. R. W. Masters, and A. Ahmed, “Quantitative analysis of btf3, hint1, ndrg1 and odc1 protein over-expression in human prostate cancer tissue,” *PLOS ONE*, vol. 8, 12 2013.
- [48] K. Cao, “Quantitative analysis of seven new prostate cancer biomarkers and the potential future of the ‘biomarker laboratory’,” *Diagnostics*, vol. 8, 07 2018.
- [49] C. Arthurs, B. N. Murtaza, C. Thomson, K. Dickens, R. Henrique, H. R. H. Patel, M. Beltran, M. Millar, C. Thrasivoulou, and A. Ahmed, “Expression of ribosomal

- proteins in normal and cancerous human prostate tissue,” *PLOS ONE*, vol. 12, pp. 1–14, 10 2017.
- [50] M. Titford, “The long history of hematoxylin,” *Biotechnic & Histochemistry*, vol. 80, no. 2, pp. 73–78, 2005.
- [51] C. Smith, “Our debt to the logwood tree: the history of hematoxylin,” *MLO: medical laboratory observer*, vol. 38, pp. 18, 20–2, 06 2006.
- [52] R. Dapson and R. Horobin, “Dyes from a twenty-first century perspective,” *Biotechnic & Histochemistry*, vol. 84, no. 4, pp. 135–137, 2009.
- [53] C. Jerónimo, P. J. Bastian, A. Bjartell, G. M. Carbone, J. W. Catto, S. J. Clark, R. Henrique, W. G. Nelson, and S. F. Shariat, “Epigenetics in prostate cancer: Biologic and clinical relevance,” *European Urology*, vol. 60, no. 4, pp. 753 – 766, 2011.
- [54] C. Jerónimo, R. Henrique, M. O. Hoque, E. Mambo, F. R. Ribeiro, G. Varzim, J. Oliveira, M. R. Teixeira, C. Lopes, and D. Sidransky, “A quantitative promoter methylation profile of prostate cancer,” *Clinical Cancer Research*, vol. 10, no. 24, pp. 8472–8478, 2004.
- [55] “Hamamatsu nanozoomer s360 digital slide scanner.” <https://www.hamamatsu.com/eu/en/product/type/C13220-01/index.html>. Accessed: 2020-09-10.
- [56] D. Wang, D. J. Foran, J. Ren, H. Zhong, I. Y. Kim, and X. Qi, “Exploring automatic prostate histopathology image gleason grading via local structure modeling,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2649–2652, 2015.
- [57] T. MUNGLE, S. TEWARY, D. DAS, I. ARUN, B. BASAK, S. AGARWAL, R. AHMED, S. CHATTERJEE, and C. CHAKRABORTY, “Mrf-ann: a machine learning approach for automated er scoring of breast cancer immunohistochemical images,” *Journal of Microscopy*, vol. 267, no. 2, pp. 117–129, 2017.
- [58] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep learning for identifying metastatic breast cancer,” 2016.

- [59] Y. Zhu, S. Zhang, W. Liu, and D. N. Metaxas, “Scalable histopathological image analysis via active learning,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, eds.), (Cham), pp. 369–376, Springer International Publishing, 2014.
- [60] M. Nalisnik, M. Amgad, S. Lee, S. Halani, J. Vega, D. Brat, D. Gutman, and L. Cooper, “Interactive phenotyping of large-scale histology imaging data with histomicsml,” *Scientific Reports*, vol. 7, 12 2017.
- [61] R. K. Padmanabhan, V. H. Somasundar, S. D. Griffith, J. Zhu, D. Samoyedny, K. S. Tan, J. Hu, X. Liao, L. Carin, S. S. Yoon, K. T. Flaherty, R. S. DiPaola, D. F. Heitjan, P. Lal, M. D. Feldman, B. Roysam, and W. M. F. Lee, “An active learning approach for rapid characterization of endothelial cells in human tumors,” *PLOS ONE*, vol. 9, pp. 1–12, 03 2014.
- [62] Z. Jia, X. Huang, E. I. Chang, and Y. Xu, “Constrained deep weak supervision for histopathology image segmentation,” *CoRR*, vol. abs/1701.00794, 2017.
- [63] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, “Weakly supervised histopathology cancer image segmentation and classification,” *Medical Image Analysis*, vol. 18, no. 3, pp. 591 – 604, 2014.
- [64] R. Sparks and A. Madabhushi, “Out-of-sample extrapolation utilizing semi-supervised manifold learning (ose-ssl): Content based image retrieval for histopathology images,” *Scientific Reports*, vol. 6, p. 27306, 06 2016.
- [65] S. Doyle, A. Madabhushi, M. Feldman, and J. Tomaszewski, “A boosting cascade for automated detection of prostate cancer from digitized histology,” vol. 9, pp. 504–11, 02 2006.
- [66] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, “Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2421–2433, 2015.

- [67] M. W. Lafarge, J. P. W. Pluim, K. A. J. Eppenhof, P. Moeskops, and M. Veta, “Domain-adversarial neural networks to address the appearance variability of histopathology images,” *CoRR*, vol. abs/1707.06183, 2017.
- [68] B. Ehteshami Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. van der Laak, “Stain specific standardization of whole-slide histopathological images,” *IEEE transactions on medical imaging*, vol. 35, 09 2015.
- [69] S. Kothari, J. Phan, and M. Wang, “Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade,” *Journal of pathology informatics*, vol. 4, p. 22, 08 2013.
- [70] M. Akbari, M. Mohrekesh, S. M. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, “Adaptive specular reflection detection and inpainting in colonoscopy video frames,” *CoRR*, vol. abs/1802.08402, 2018.
- [71] E. Blotta, A. Bouchet, V. Ballarin, and J. Pastore, “Enhancement of medical images in HSI color space,” *Journal of Physics: Conference Series*, vol. 332, p. 012041, dec 2011.
- [72] Ming-Kuei Hu, “Visual pattern recognition by moment invariants,” *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.
- [73] P. Yap, X. Jiang, and A. Chichung 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, 2010.
- [74] M. R. Teague, “Image analysis via the general theory of moments\*,” *J. Opt. Soc. Am.*, vol. 70, pp. 920–930, Aug 1980.
- [75] M. Liu, Y. He, and B. Ye, “Image zernike moments shape feature evaluation based on image reconstruction,” *Geo-spatial Information Science*, vol. 10, no. 3, pp. 191–195, 2007.
- [76] D. Zhang and G. Lu, “Lu, g.: Review of shape representation and description techniques. pattern recognition 37, 1-19,” *Pattern Recognition*, vol. 37, pp. 1–19, 01 2004.

- [77] C. . Teh and R. T. Chin, “On image analysis by the methods of moments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 496–513, 1988.
- [78] 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.
- [79] C. Singh and P. Sharma, “Local and global features based image retrieval system using orthogonal radial moments,” *Optics and Lasers in Engineering*, vol. 50, pp. 655–667, 10 2012.
- [80] P. Sharma, “Performance analysis of various local and global shape descriptors for image retrieval,” *Multimedia Systems*, vol. 19, pp. 339–357, 07 2013.
- [81] Q. Chaudry, S. Raza, A. Young, and M. Wang, “Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features,” *Signal Processing Systems*, vol. 55, pp. 15–23, 04 2009.
- [82] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, J. Saltz, and M. Gurcan, “Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading,” *Journal of Signal Processing Systems*, vol. 55, pp. 169–183, 04 2009.
- [83] M. R. K. Mookiah, M. Pal, R. Paul, C. Chakraborty, J. Chatterjee, and A. Ray, “Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis,” *Journal of medical systems*, vol. 36, pp. 1745–56, 12 2010.
- [84] J. Iglesias-Rozas and N. Hopf, “Histological heterogeneity of human glioblastomas investigated with an unsupervised neural network (som),” *Histology and histopathology*, vol. 20, pp. 351–6, 05 2005.
- [85] B. Lessmann, T. Nattkemper, V. Hans, and A. Degenhard, “A method for linking computed image features to histological semantics in neuropathology,” *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 631 – 641, 2007. Intelligent Data Analysis in Biomedicine.

- [86] I. M. Stephanakis, G. C. Anastassopoulos, and L. S. Iliadis, “Color segmentation using self-organizing feature maps (sofms) defined upon color and spatial image space,” in *Artificial Neural Networks – ICANN 2010* (K. Diamantaros, W. Duch, and L. S. Iliadis, eds.), (Berlin, Heidelberg), pp. 500–510, Springer Berlin Heidelberg, 2010.
- [87] M. Datar, D. Padfield, and H. Cline, “Color and texture based segmentation of molecular pathology images using hsoms,” in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 292–295, 2008.
- [88] L. A. D. Cooper, J. Kong, D. A. Gutman, F. Wang, S. R. Cholleti, T. C. Pan, P. M. Widener, A. Sharma, T. Mikkelsen, A. E. Flanders, D. L. Rubin, E. G. V. Meir, T. M. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, “An integrative approach for in silico glioma research,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2617–2621, 2010.
- [89] H. Chang, G. Fontenay, J. Han, G. Cong, F. Baehner, J. Gray, P. Spellman, and B. Parvin, “Morphometric analysis of tcga glioblastoma multiforme,” *BMC bioinformatics*, vol. 12, p. 484, 12 2011.
- [90] T. Kohonen, M. R. Schroeder, and T. S. Huang, *Self-Organizing Maps*. Berlin, Heidelberg: Springer-Verlag, 3rd ed., 2001.
- [91] J. Zurada, *Introduction to Artificial Neural Systems*. USA: West Publishing Co., 1992.
- [92] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium, “Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer,” *JAMA*, vol. 318, pp. 2199–2210, 12 2017.
- [93] R. E. Thomson and W. J. Emery, “Chapter 4 - the spatial analyses of data fields,” in *Data Analysis Methods in Physical Oceanography (Third Edition)* (W. J. Emery and R. E. Thomson, eds.), pp. 313 – 424, Boston: Elsevier, third edition ed., 2014.



- [94] Y.-S. Park, J. Tison, S. Lek, J.-L. Giraudel, M. Coste, and F. Delmas, “Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across france,” *Ecological Informatics*, vol. 1, no. 3, pp. 247 – 257, 2006. 4th International Conference on Ecological Informatics.

# Appendix

## Summary of interview with Rui and Dr Aamir

### Questionnaire

1. Brief introduction: About your work, experience and years dedicated, tissue management and classification expertise.
2. Tell us more details on disease severity and current diagnosis techniques (including accuracy).
3. What is the minimum image quality required to be classified correctly? Does it matter on type of image?
4. What exactly do you look for in image? Boundary? Shape? Color?
5. Please describe naming convention of image shared (ex: JLTA1\_x40\_z0\_A1\_T)? More specifically, details on slides and cores table used (going from A to J and 1 to 6) in naming.

### Recorded minutes of meeting

- (Aamir) Manual Analysis is mostly intuitive due to years of practice of looking at H&E images.
- (Rui) Pathologists usually use 40x zoom under microscope and sometimes require up to 400x to be certain.
- (Aamir) Sometimes (~5%) Immunohistochemistry (IHC) is used in diagnosis to confirm borderline cases. However, it's not always accurate.

1. Works due to the presence (or absence) of basal cells, detected by specific antibodies against it combined with racemase expression in luminal epithelial cells.
- (Aamir & Rui) A brief discussion on embryonal carcinoma and yolk sac.
  - (Rui) Prostate cancer diagnosis techniques in H&E images are:
    1. Looking for small glands in front of large glands.
    2. Haphazard distribution.
    3. Higher density of nuclei and larger nuclei (nucleomegaly)
    4. Presence of prominent nucleoli.
  - (Aamir & Rui )Gleason score for grading prostate cancer.
    1. This score is based on how much the cancer looks similar to small benign glands when viewed under a microscope.
    2. Based on how the cancer cells are arranged a score on a scale of 3 to 5 is assigned (patterns); due to morphological heterogeneity, in many cases more than one pattern is present. Depending on the proportion, predominance and potential aggressiveness, as well as whether grading is performed on prostate biopsy or surgical specimen, the final grade combines two figures: e.g. 3+4=7 (grade group 2); if only one pattern present, then, the pattern is doubled: e.g., 3+3=6 (grade group 1)
  - (Rui) In H&E staining Hematoxylin colors the nuclei of cells blue or dark-purple and Eosin stains the cytoplasm and some other structures including extracellular matrix such as collagen in up to five shades of pink. In cancer glands there is usually a blue-tinged secretion in the lumen and, sometimes, pink crystal-like structures (crystalloids)

### **Approval to use prostate tissue images shared by Dr Aamir**

Consent: 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. Institutional Review Board approved their use for research purposes. Only anonymized images from hematoxylin and eosin stained slides were used in this project.

We have the REC (ethics committee approval in the UK) to use the tissue samples. You did not use any 'human tissue' in your study, only images, so that does not apply.