DEVELOPMENT OF PATTERN ANALYSIS AND MACHINE LEARNING TECHNIQUE FOR CANCER DIAGNOSIS

A thesis submitted to the
University of Petroleum and Energy Studies

For the Award of
Doctor of Philosophy
in
Computer Science and Engineering

By
ANIL KUMAR

December 2020

Supervisor
Dr. Manish Prateek

School of Computer Science
University of Petroleum and Energy Studies
Energy Acres, P.O. Bidholi via Prem Nagar,
Dehradun, 248007: Uttarakhand, India.

DEVELOPMENT OF PATTERN ANALYSIS AND MACHINE LEARNING TECHNIQUE FOR CANCER DIAGNOSIS

A thesis submitted to the
University of Petroleum and Energy Studies

For the Award of
Doctor of Philosophy
in
Computer Science and Engineering

By
ANIL KUMAR
(SAP ID 500024317)

December 2020

Supervisor
Dr. Manish Prateek
Professor,
School of Computer Science
University of Petroleum & Energy Studies

School of Computer Science
University of Petroleum and Energy Studies
Energy Acres, P.O. Bidholi, via Prem Nagar,
Dehradun, 248007: Uttarakhand, India

Declaration

I declare that the thesis entitled **“Development of Pattern Analysis and Machine Learning Technique for Cancer Diagnosis”** has been prepared by me under the guidance of Dr. Manish Prateek, Professor at School of Computer Science, University of Petroleum & Energy Studies. No part of this thesis has formed the basis for the award of any degree or fellowship previously.

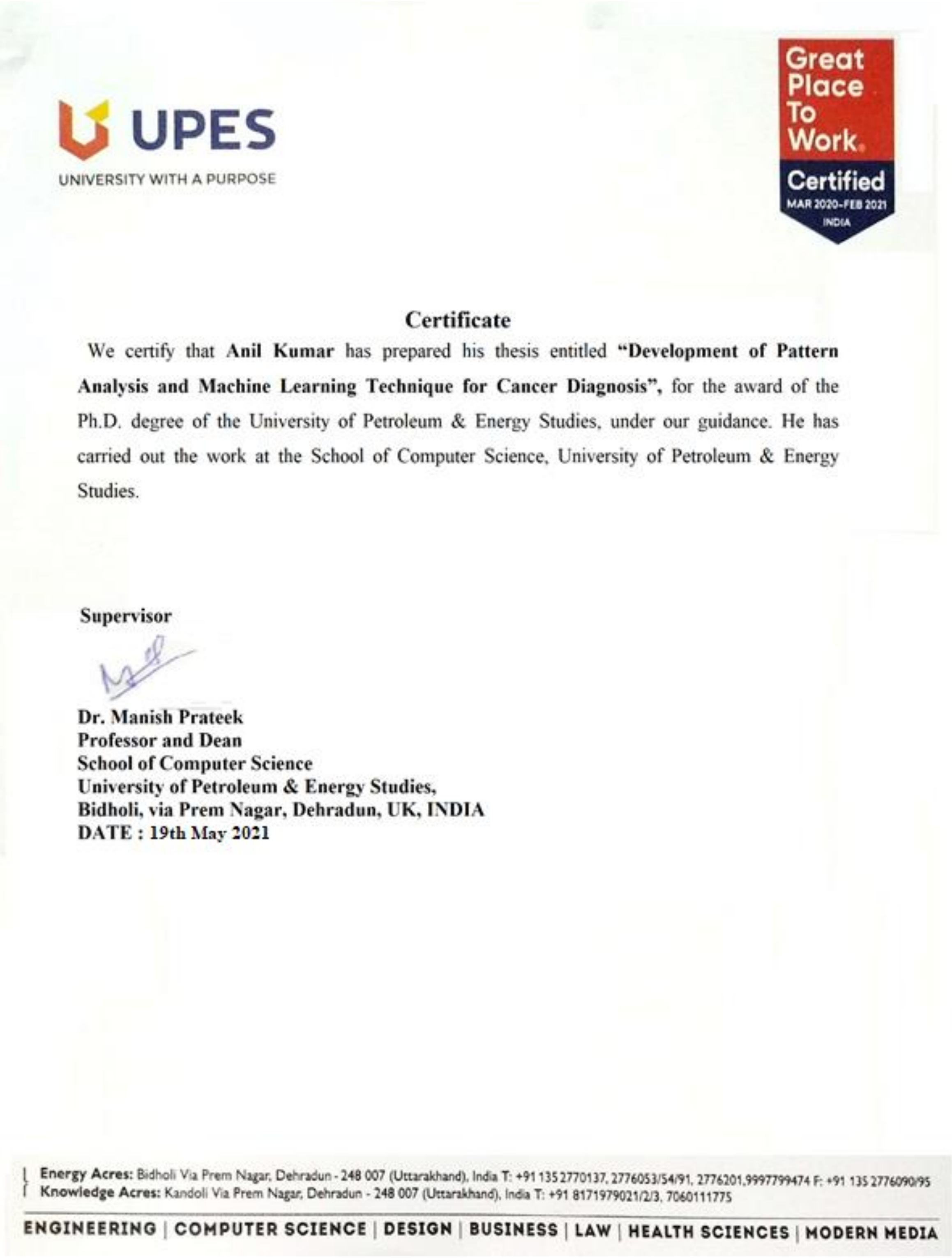
Anil kumar

ANIL KUMAR
School of Computer Science

University of Petroleum & Energy Studies,

Bidholi via Prem Nagar, Dehradun, UK, INDIA

DATE: 19th May 2021



UPES
UNIVERSITY WITH A PURPOSE

Great Place To Work. Certified
MAR 2020-FEB 2021
INDIA

Certificate

We certify that **Anil Kumar** has prepared his thesis entitled **"Development of Pattern Analysis and Machine Learning Technique for Cancer Diagnosis",** for the award of the Ph.D. degree of the University of Petroleum & Energy Studies, under our guidance. He has carried out the work at the School of Computer Science, University of Petroleum & Energy Studies.

Supervisor

Dr. Manish Prateek
Professor and Dean
School of Computer Science
University of Petroleum & Energy Studies,
Bidholi, via Prem Nagar, Dehradun, UK, INDIA
DATE : 19th May 2021

Energy Acres: Bidholi Via Prem Nagar, Dehradun - 248 007 (Uttarakhand), India T: +91 135 2770137, 2776053/54/91, 2776201,9997799474 F: +91 135 2776090/95
Knowledge Acres: Kandoli Via Prem Nagar, Dehradun - 248 007 (Uttarakhand), India T: +91 8171979021/2/3, 7060111775

ENGINEERING | COMPUTER SCIENCE | DESIGN | BUSINESS | LAW | HEALTH SCIENCES | MODERN MEDIA

Abstract

Breast cancer is one of the major modern day problem in medical science that causes the death of thousands of women worldwide. One of the effective measures to confront this disease is early detection using an electron microscope generally applied by pathologists and clinicians on biopsy slides. The human intervention process of exposition of biopsy slides imply the identification of breast cut (e.g., masses part of biopsy), the segmentation of sample breast cut boundaries, the classification of it based on their color after staining, structure, emergence, and textural features. The manual analysis of breast cut from biopsy slides represents a large exposition variability amongst pathologists. This kind of problems and mutability can be reduced with the adaptation of the technique called computer aided diagnosis (CAD) that can assist the pathologists for analysis of breast lesions (cut) with clarity and high precisions. However, the CAD system is very useful in a clinical labs that efficiently help to clinicians to classify the breast lesions as malignant or benign as the extended scope of the work.

Digital pathology is strongly evolved in the last two decades and training for the changes and development is required for the pathologists. It includes both the imaging of tissues and remote consultations, also called telepathology, widely used as an option for a second opinion. To support the pathologists and clinicians to make a more comfortable and accurate diagnosis of the biopsy slides, it is developed the digital Whole Slide Imaging (WSI) system. Automated identification of the hot spots, image segmentation, and the classification of breast lesions are majorly the three steps involved in advanced CAD to analyze the WSI samples. Due to a large size of the whole slide images, it is difficult to display, read, process, identify and localize the region of interests (ROIs), and archives the digital slides.

The machine learning algorithms have a vital role to implement the clinical decision support system (CDSS) used for pattern recognition and classifications. The CDSS is an integrated part of CAD. The WSI is supported by fluorescence, immunohistochemistry (IHC), and multispectral imaging concepts. It is observed the computational challenges because of the complexity and large size of WSI sample. The goal of the research work is to identify and localize the ROIs on WSI slides and try to grade the breast cancer based on the proliferation score. The Ki-67 antigen is one of the suitable biomarkers to identify and differentiate between the immunopositive and immunonegative cells. The unsupervised machine learning algorithm is supported by shape formulas and morphological features implemented on ICIAR 2018 BACH datasets for finding and localizing the ROIs. The achieved accuracy of the work is 85.5% using the similarity measure formula intersection over union (IoU). The accuracy is better than many existing state-of-the-art machine learning and deep learning algorithms.

Grading of cancer is achieved by counting of immunopositive and immunonegative nuclear sections in a sample that determines the proliferation score. The BreCaHAD dataset contains a variety of malignant cases of different patients. It provides 40x magnification H&E stained microscopic histopathology images saved in .tiff format with RGB band. The procedure begins with preprocessing, which focuses on resizing, smoothing, and enhancement in this study. After preprocessing, the RGB sample is decomposed into the HSI color space. The BreCaHAD data set is stained with H&E, with brown and blue colour levels playing a key role in distinguishing immunopositive from immunonegative nuclear sections. A natural trait of H&E Ki-67 is the Blue color in RGB and the Hue in HSI color space. After segmentation, it is using Otsu's thresholding and unsupervised machine learning to compute the shape parameters. The morphological operators aid in the resolution of the problem of overlapping nuclear sections in sample images, allowing for accurate counting and automatic segmentation.

It is effectively possible to predict the label or grade of breast cancer using major morphological features and an unsupervised machine learning technique on the BreCaHAD dataset. The performance measures like accuracy: 90.8%, f-score: 94.74%, precision: 95.7%, recall: 93.8%, specificity: 0.6803, Balance Classification Rate (BCR): 0.7975 and Mathew's Correlation Coefficient

(MCC): 0.5855 are obtained in proposed methodology which is better than existing techniques.

Acknowledgment

I am highly indebted to my guide **Dr. Manish Prateek** who has guided me all through the completion of this Ph.D thesis. My enthusiasm in this subject was inspired by their profound knowledge in this area and elegant research style. They have helped me with wise advice, useful discussions, comments, and facilities. They never hesitated to spend precious time and effort to guide my work. My guides helped me to overcome the various difficulties and challenges that were raised at various stages of the project. It is just ineffable to express my deep gratitude towards them.

My special thanks to Dr. Himanshu Madhav who guided me and shown me the path for this interdisciplinary research work. I am very grateful to Dr. Neena Chauhan, Dr. S. K. Verma (Cancer Research Institute, Himalayan Hospital, Jolly Grant, Dehradun, India) to give an understanding of tissue biopsy and other IHC steps in their labs. I am also grateful to my friends Dr. Ravi Tomar, Ms. Poonam Kainthura and Mr. Gagan Deep Singh, SoCS, for their enlightened guidance and encouragement, and Mr. Rajat Garg, JRF(NISAR Project at UPES) for their support during the preparation of the thesis.

I acknowledge my indebtedness to Dean (R&D) Dr. Devesh Kumar Avasthi, Ex. Associate Dean (R&D) Dr. J. K. Pandey, Assistant Dean (R&D) Dr. Kiran Kumar Ravulakollu for their all time support, Associate Librarian Dr. Prem Prakash Sati, Mr. Jatendra Sharma Sr. Manager, and Dharmendra Chauhan, Department of IT, University of Petroleum & Energy Studies Dehradun for proving me the library and computing resources as and when required.

I would like to apologize to those whose names do not figure here but have helped me during the tenure of my research. Lastly, I would like to dedicate this work to my parents, and other family members for being present with me all the time with their blessings.

ANIL KUMAR

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
ALEXNET	Alex Krizhevsky Network
ASCO	American Society of Clinical Oncology
ASCO	American Society of Clinical Oncology
AUC	Area Under Curve
BACH	Breast Cancer Histology Images
BCR	Balance Classification Rate
BreCaHAD	Breast Cancer Histopathological Annotation and Diagnosis
CAD	Computer Aided Diagnosis
CCD	Charged Couple Device
CDSS	Clinical Decision Support System
CMOS	Complementary Metal Oxide Semiconductor
CMYK	Cyan Magenta Yellow and Key
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
DCNN	Deep Convolutional Neural Network
DCO	Distal Clavicle Osteolysis
DIA	Digital Image Analysis
DICOM	Digital Imaging and Communications in Medicine
DLBP	Dominant Local Binary Pattern
EM	Expectation-Maximization
ER	Estrogen Receptor
FC	Fully Connected
FDA	Food & Drug Administration
FN	False Negative
FNAB	Fine Needle Aspiration Biopsy
FP	False Positive
FROC	Free-Response Operating Characteristic
GB	Giga Byte
GLCM	Gray Level Co-occurrence Matrix
GLOBOCAN	Global Cancer Observatory of Cancer
GMM	Gaussian Mixture Model
H&E	Hematoxylin and Eosin
HDI	High Definition Intensity
HDI	Human Development Index
HER2	Human Epidermal growth factor Receptor 2
HIS	Hue Saturation Intensity
HREBA	Health Research Ethics Board of Alberta
ICIAR	International Conference

ICMR	Indian Council of Medical Research
IDC	Invasive Ductal Carcinoma
IHC	immunohistochemistry
IoU	Intersection Over Union
KNN	k – Nearest Neighbor
KRAS	Kirsten Rat Sarcoma
MCC	Mathew’s Correlation Coefficient
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NCDIR	National Centre for Disease Informatics and Research
NCG	National Cancer Group
NCRP	National Cancer Registry Programme
NIH	National Institute of Health
NLD	Nuclear section Longest Diameter
NSD	Nuclear section Shortest Diameter
OOF	Out of Focus
OPOD	One Patch in One Decision
PACS	Picture Archive and Communication System
PBCR	Population-Based Cancer Registries
PC	Personal Computer
PR	Progesterone Receptor
PS	Proliferation Score
RAM	Random Access Memory
RGB	Red Green Blue
RMSE	Root Mean Square Error
ROI	Region of Interest
SPM	Spatial Pyramid Matching
SSIM	Structural Similarity Index
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TIFF	Tagged Image File Format
TN	True Negative
TP	True Positive
TUPAC	Tumor Proliferation Assessment Challenge
USFDA	United States Food and Drug Administration
VM	Virtual Microscopy
WHO	World Health Organization
WSI	Whole Slide Imaging

List of Figures

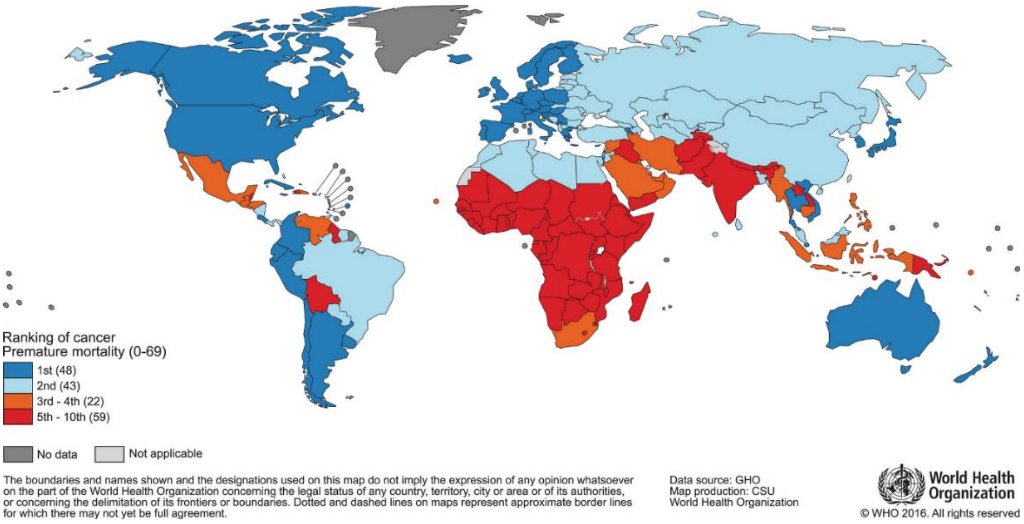
List of Tables

Chapter 1

Introduction and Motivation

1.1 Cancer Statistics

Cancer has always affected human life. Until recently, the global impact and dispersion of cancer were only known in a few countries and communities. But nowadays, due to the technological enhancement, the information has been expanded globally and has a reasonable basis for the estimation of cancer data globally. One of the main reason of premature casualty is cancer to those aged 30 to 69 in 134 out of 183 nations, and it ranks 3rd or 4th in the remaining 45 countries. A total of 15200,000 early deaths were registered from noncontagious illness worldwide in 2016 and 4.5 million about 29.8% were due to the various cancer types. Top six most general carcinomas found worldwide are lung cancer, cervical cancer, stomach cancer, breast cancer, prostate cancer, and colorectal cancer with global trends in cancer occurrence and death [1]. Lung cancer is the regular cancer in the world in terms of both occurrence (2.1 million new instances in 2018) and death (18000000 deaths in 2018). Smoking Tobacco is the strong reason of lung cancer, accounting for 63 percent of overall fatalities. More than 90 percent of lung cancer casualties in countries where both men and women smokes. The 3rd leading cancer type is colorectal cancer throughout the world, affecting both men and women (1800000 new cases in 2018). In terms of casualties, it is ranked second (880000 casualties in 2018). The data shows that the death is much lower than occurrences reflects the most cases have a positive prognosis. The countries which have high HDI, relatively have high trends of colorectal cancer [2]. In women, breast cancinoma is the most typical diagnosed cancer. The reported diagnosis was 2100000 new cases in 2018. In 2018, the major cause of cancer death in women was estimated to be over 627000 deaths worldwide. Globally, prostate cancer is diagnosed as men's second most common cancer. In 2018, 1300000 new cases were estimated and total 13.5 percent of all new cancer cases were found in male. It is a fewer common cause of cancer-related death, accounting for 360000 (6.7% of casualties due to cancer in men) in 2018 [3]. The 5th most common cancer type is stomach cancer worldwide, with an expected 1000000 new cases (5.7 percent of all new cancer cases) in 2018. Although, it ranks third in terms of fatality due to its dismal prognosis (783 000 deaths in 2018). With an anticipated 570000 new cases and 311000 casualties in 2018. The fourth most frequent cancer type is cervical cancer happened in women globally in terms of both occurrence and casualties [4]. The global map is representing the different zones and the respective national ranking of death due to cancer at ages below 70 Years in 2015 is appeard in Figure 1.1.



The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data source: GHO
Map production: CSU
World Health Organization

© WHO 2016. All rights reserved

Figure 1.1: The World Health Organization (WHO) published a global map depicting the national rankings of cancer as a cause of death in people under the age of 70 in 2015

1.1.1 Global and Local Scenario

In emerging countries like India, cancer is projected to be a serious issue. The "GLOBOCAN (An International Agency for Research on Cancer)" accounted with a projected occurrence of 1000000 cases in 2012 and 1700000 cases in 2035. It is anticipated that the India's cancer growth to nearly quadruple. In addition, cancer fatalities are anticipated to increase from 680,000 in 2012 to 1.2 million in 2035 [5]. According to the GLOBOCON 2018, it is calculated 18000000 new cancer cases (excluding nonmelanoma skin cancer estimated as

17000000) and 9600000 casualties due to cancer each year (excluded 9500000 cases of nonmelanoma skin cancer). The most general diagnosed category of cancer is lung cancer (11.6 percent of all cases) and the major cause of casualties due to cancer (18.4 percent of all cancer deaths) in both men and women, closely followed by female breast cancer (11.6 percent), colorectal cancer (6.1 percent), prostate cancer (7.1 percent), and incidence stomach cancer (8.2 percent) and colorectal cancer (9.2 percent). Lung carcinoma is the most common cancer among men and the primary of reason of casualty. The next type of cancer caused for death is the prostate and colorectal cancer then liver and stomach cancer. The most general diagnosed cancer among women is breast cancer, and it is also one of the leading cause of casualty due to cancer. Colorectal and lung cancers steps second and third, respectively, in terms of incidence and death. The fourth most general malignancy in both ways, casualty and incidence is cervical cancer [4].

India's mortality to incidence ratio of 0.68 is significantly greater than that of high HDI countries (0.57) and very high HDI countries (0.38). Although few of this discrepancy might be because of over diagnosis due to screening in developed countries, part of it is because of inequitable distribution and inaccessibility of healthcare resources to vast parts of the country [6]. In underdeveloped nations, lower survival rates are likely attributable to a combination of advanced stage diagnosis, limited access to high-quality cancer care, and patients' inability to finance the best treatment options. These issues need to be addressed on multiple fronts: Patients and primary care physicians need to be more aware of cancer as a curable disease, making cancer treatment accessible to individuals at their homes, and discovering novel, cost-effective diagnostic and treatment methods.

There are differences in the occurrence and casualty due to the cancer in different regions of Republic India. The standards of cancer diagnosis and prognosis may differ considerably between institutions, states, and geographical subregions [7]. There is a lack of uniformity in criteria set for prevention, early diagnosis, evidence-based treatment, and follow-up of patients with cancer. This disparity has manifested primarily because of a lack of an established network of cancer centers across the country to implement common standard management guidelines. Though regional cancer centers exist in all parts of the country and geographically cover the population, they too have varying standards of care. One of the biggest challenges and needs for effective cancer control in India is for uniformly high standards of care to be provided throughout the country. Dissemination of high-quality cancer care across the country and ensuring uniformity of standards would eliminate the need for patients to have to travel long distances for optimal medical care.

To war against cancer Government of India had constituted the “National Cancer Grid” in Aug 2012 with the directive of connected cancer centers across India. It began as a small endeavor with 14 cancer centers, but it has quickly grown to encompass 52 major cancer centers across the country, making it one of the world's largest cancer networks. The NCG, which is funded by the Indian government's Department of Atomic Energy, has the primary goal of achieving uniform standards of care across India by implementing evidence-based management practices across these institutions. It also aims to ease the sharing of expertise among centers and to establish a ready network of cancer research centers [8].

The global impact of cancer is also very harsh. It was surveyed that 14.1 million new cancer instances found worldwide in 2012. In 2008, cancer claimed the lives of 169.3 million healthy people around the world. By 2030, there will be 23.6 million new cancer instances every year over the world (estimated). According to Cancer Research Manchester (UK), more than four out of ten cancers originate in nations with a low or medium Human Development Index (HDI).

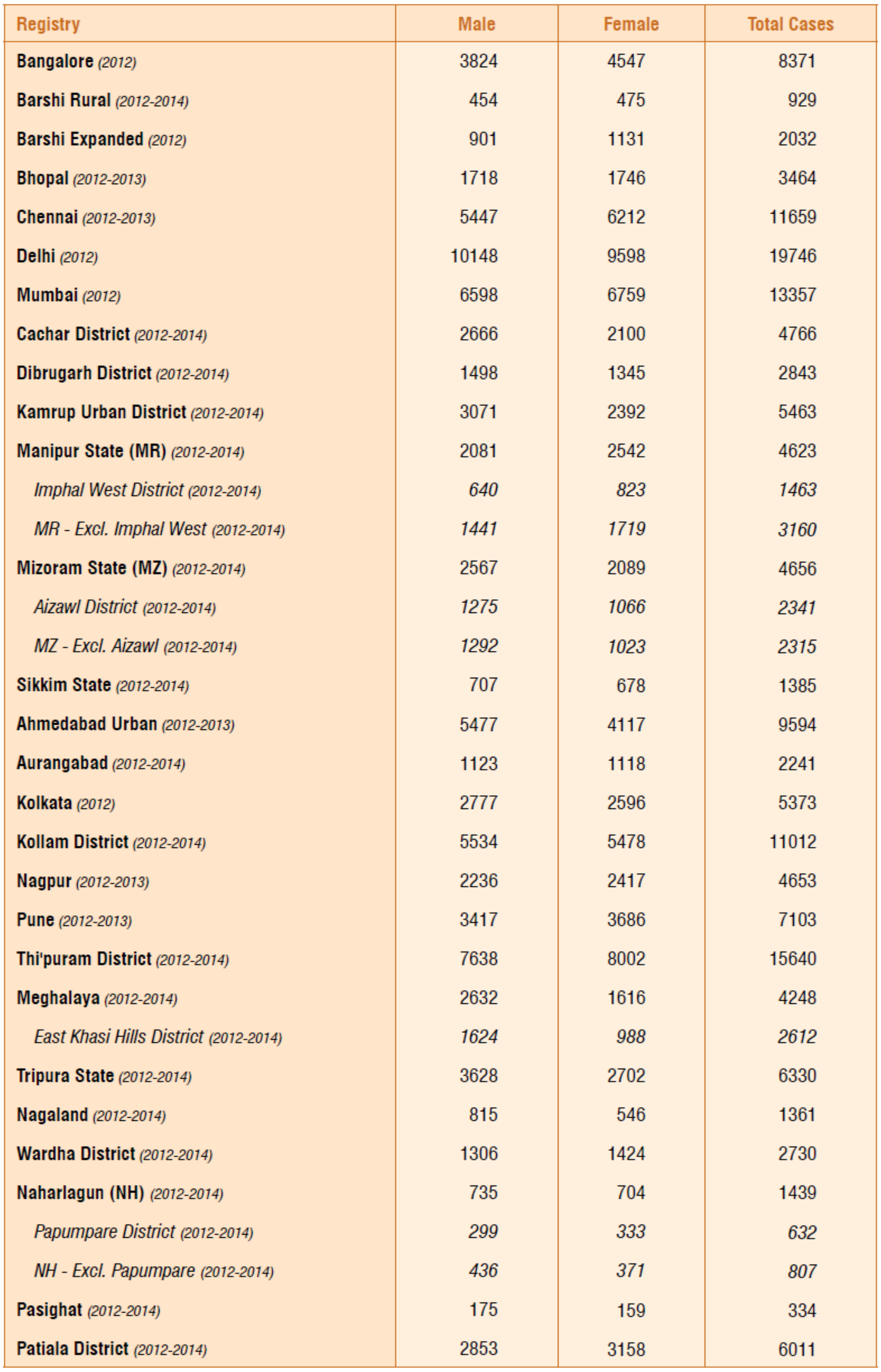
Based upon the World Cancer Report, the most thorough global analysis of the disease to date, cancer growth rate could rise to 15 million new cases by 2020. Central and South America, Africa, and Asia region count for more than 60% of total new cases each year globally. These areas account for 70% of all cancer fatalities worldwide. Cancer appears to be strengthening its grip on India, with a million new cases reported each year. According to experts, the fatal disease's occurrence in India is anticipated to increase five-fold by 2025 [6].

To address challenges of turn-round time and costs in Pathology Operations, the project would offer customized, integrated & advanced technological products & solutions. Integrated Digital slide imaging and management would be the niche offering (in Whole Slide Imaging Segment), making it easy for hospitals, Research Organizations, Educations Institutes, etc. to efficiently and effectively manage cancer diagnostic cases. The solutions aim at enabling ease of access for Pathologists to analyze slides with Clarity & Precision.

Every year, it is diagnosed around 1000000 new cases of cancer in India, with a total occurrences of 2.5 million. Cancer is responsible for 6% of adult fatalities in the country, out of a total of 700000 deaths per year. Among various diseases, cancer has become one of the enormous threats to our society. According to census data of India, the rate of mortality due to cancer was high and a worrying situation, with around 806000 cases reported by the turn of the century. With nearly 0.3 million fatalities each year, cancer is India's 2nd most common disease and one of the leading cause of casualty. This is due to the disease's lack of prevention, diagnosis, and treatment options. The skin, lungs, breast, stomach, esophagus, prostate, liver, cervix, rectum, bladder, blood, mouth, and other organs cancer subtypes have been found among Indian population. The major reason of evolution of such diseases categorized as Internal (hormonal, inadequate immune conditions, genetic, mutations,) and external (population overgrowth, food habits, industrialization, social, etc.) variables may be to blame for the high incidence rates of various malignancies [6].

In India, the “National Centre for Disease Informatics and Research (NCDIR) - National Cancer Registry Programme (NCRP), Bengaluru (India)”, part of the “Indian Council of Medical Research (ICMR)”, presented a three-year study on 27 “Population-Based Cancer Registries (PBCR)” from 2012 to 2014 as listed in Table 1.1. Individual core data was provided by PBCR. The NCDIR-NCPR in Bengaluru, India, performed quality control checks, tabulations, and statistical analysis. The report is about the information on cancer incidence in India.

Table 1.1: Total number of cases registered for all 27 PBCRs provides information about 34 geographical areas



Registry	Male	Female	Total Cases
Bangalore *(2012)*	3824	4547	8371
Barshi Rural *(2012-2014)*	454	475	929
Barshi Expanded *(2012)*	901	1131	2032
Bhopal *(2012-2013)*	1718	1746	3464
Chennai *(2012-2013)*	5447	6212	11659
Delhi *(2012)*	10148	9598	19746
Mumbai *(2012)*	6598	6759	13357
Cachar District *(2012-2014)*	2666	2100	4766
Dibrugarh District *(2012-2014)*	1498	1345	2843
Kamrup Urban District *(2012-2014)*	3071	2392	5463
Manipur State (MR) *(2012-2014)*	2081	2542	4623
Imphal West District (2012-2014)	*640*	*823*	*1463*
MR - Excl. Imphal West (2012-2014)	*1441*	*1719*	*3160*
Mizoram State (MZ) *(2012-2014)*	2567	2089	4656
Aizawl District (2012-2014)	*1275*	*1066*	*2341*
MZ - Excl. Aizawl (2012-2014)	*1292*	*1023*	*2315*
Sikkim State *(2012-2014)*	707	678	1385
Ahmedabad Urban *(2012-2013)*	5477	4117	9594
Aurangabad *(2012-2014)*	1123	1118	2241
Kolkata *(2012)*	2777	2596	5373
Kollam District *(2012-2014)*	5534	5478	11012
Nagpur *(2012-2013)*	2236	2417	4653
Pune *(2012-2013)*	3417	3686	7103
Thi'puram District *(2012-2014)*	7638	8002	15640
Meghalaya *(2012-2014)*	2632	1616	4248
East Khasi Hills District (2012-2014)	*1624*	*988*	*2612*
Tripura State *(2012-2014)*	3628	2702	6330
Nagaland *(2012-2014)*	815	546	1361
Wardha District *(2012-2014)*	1306	1424	2730
Naharlagun (NH) *(2012-2014)*	735	704	1439
Papumpare District (2012-2014)	*299*	*333*	*632*
NH - Excl. Papumpare (2012-2014)	*436*	*371*	*807*
Pasighat *(2012-2014)*	175	159	334
Patiala District *(2012-2014)*	2853	3158	6011

According to all the 27 PBCRs pooled data records, it is shown that ten leading sites of cancer, city wise for males (2012-14) in Table 1.2 and the same for women in Table 1.3 [7]. The relative proportion (%) of cancers based on different methods of diagnosis, considered as most valid are represented in Figure 1.2. The data says that for 85.3% male and 86.9% female patients, the medium of diagnosis is imaging and microscopic concepts. The choices for rest are DCO, Clinical, X-Ray, and others [8].

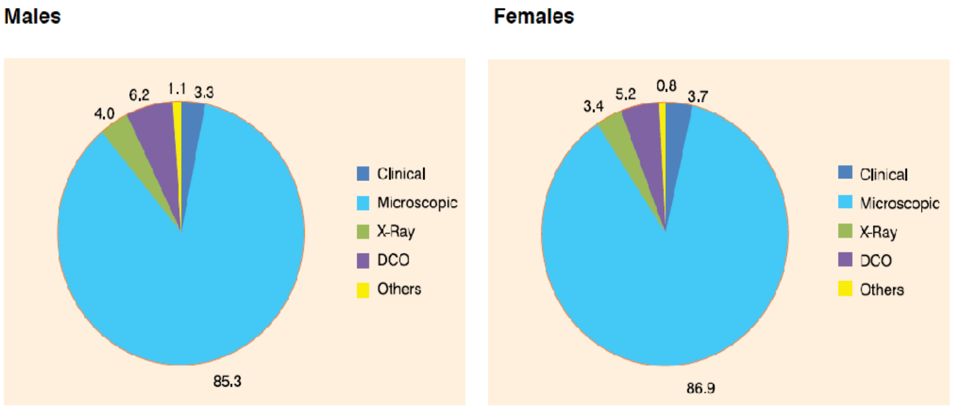


Figure 1.2: Relative Proportion (%) of Cancers Based on Different Methods of Diagnosis – All PBCRs (Pooled Data)

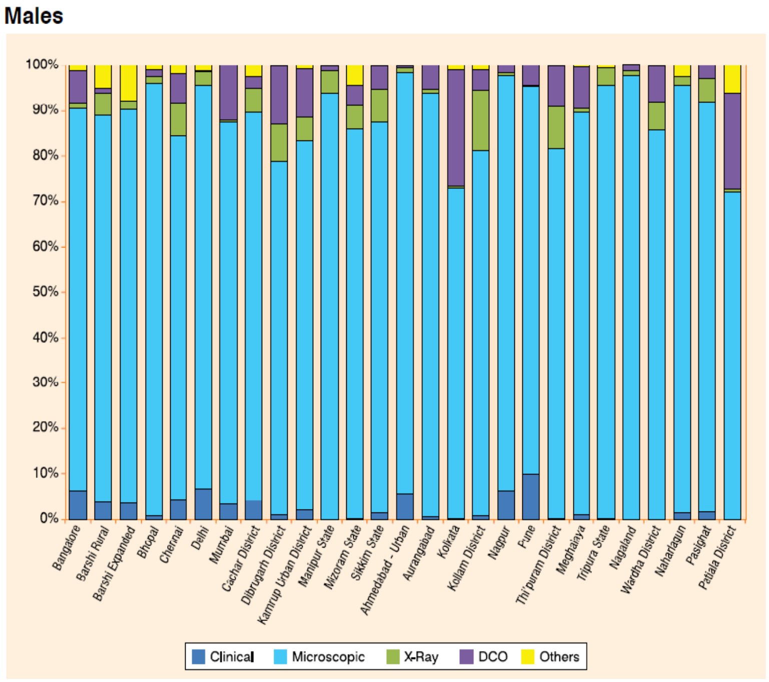


Figure 1.3: Proportion (%) of Cancers based on Different Methods of Diagnosis - All PBCRs for Males

Table 1.2: Ten Leading Sites of Cancer City Wise for Males (2012-14)

City	Lung	Stomach	Prostate	Oesophagus	Brain, NS	NHL	Liver	Tongue	Mouth	Colon	Rectum	Larnyx	Hypopharynx	Penis	Urinary Bladder	Gall Bladder	Lymphoid Leuk.	Tonsil	Nasopharynx	Other Skin	Myeloid Leuk.	Pharynx Unsp.	Leuk. Uns	Pancreas
Banglore	10.6	7.6	7	6	5	4.4	4.3	4.2	3.9	3.6														
Barshi Rural		4.2	4.4	7.7		4.4	6.2	4.4	8.6		4.4	4.2	4.2											
Barshi Expanded	5	4.1	4.1	6.7		3.7	4.3	8.1	9			3.4		3.4										
Bhopal	11.3		4.8	4.4	2.7	3		9.1	15.3			5.8	4.4		2.7									
Chennai	9.8	9.3	5.4	3.9		3.7	3.9	7	8	3.7		4.2												
Delhi	10.5		6.7	4		4.4		6.5	6.9			5.7			4.4	3.6	3.5							
Mumbai	10.4	4.3	7.9	3.7	3.9	3.7	5.7	4.9	8.8			4.1												
Cachar District	8.4	4.2		10		2.1		4.6	5.7			5.4	8.6			4.2		1.9						
Dibrugarh District	5.1	7.9		15			3.7	4.9	6.8	3.9	2.8		12			4.5								
Kamrup Urban District	8.4	6.7	4.8	14			3.3	3.6	5.2			3.6	8.9			4.2								
Manipur State (MR)	17.2	7.1		5.2		4.7	4.1				3.4	3.7							5.9	4.5	3.2			
Imphal West District	16.6	4.1		6.4			5.6	3.1		4.4	3.3	3.4								5.9	3.4			
MR-Excl. Imphal West	17.5	8.4		4.7		5.5	3.4				3.4	3.8							7.3	3.9	3.1			
Mizoram State (MZ)	14.1	19		16			5.5	1.9		2.5	3.4	2.8	5						2.5					
Aizwal District	12.6	16		19		1.9	5.2	2.7		2.4	3.5	2.8	6.5											
MZ-Excl.Aizwal	15.6	21	1.4	13			5.7			2.6	3.3	2.9	3.6						3.3					
Sikkim State	7.2	16		6.8	3		7.6		4.7			3.8							5.4	4		3.5		
Ahmedabad Urban	8.4		3.7	5.6		2.6		12	20.3	2.5		3.6	3.3								2.8			
Aurangabad	12.3	3.5		6.8		3.7		9.6	11.8		3.8	6.6	3.1				4.4							
Kolkata	18.9	3.5	8.2				3	5.5	6.9	4.1		5.4			3.9	3.3								
Kollam District	18.2	4.7	5.2	3.4		3.9	4.7	4.3	5.6		4.3	4.5												
Nagpur	6.8	3.4		7	3.1	2.9	2.8	8	15.7		3.2	6												
Pune	8.3		7.3	4.6	4.2	4.7	4.5	5.7	10.6		3.9	4.4												
Thi'puram District	13	3.8	7.2			3.5	4	5.1	5.1		4.2	4.2			3.8									
Meghalaya	6.4	6.8		31			2.2	5.7	4.5			4.7	9					4.3				1.9		
East Khasi Hills District	5.5	5.9		34			2.2	5.5	5			5.1	10					4.6				1.7		
Tripura State	17.6	6.1		7.9			2.6	5.5	6			6.6	5.7			3.3		2.7						
Nagaland	6.1	12		10			2.9		4.2	2.6		5.9	5.9					3.3	15					
Wardha District	6.4		3.2	7	5	3.5	6.2	5.2	14.6											4.1	3.2			
Naharlagun (NH)	6.8	25		7.8			20		1.9			3.3	2.2						4.4	3.4			3	
Papumpare District	9.7	21		11			14		3.7			4	3						4.4	5			3	
NH-Excl. Papumpare	4.8	27		5.7			24	2.1				2.8							4.4	2.3			4	2.3
Pasighat	4.6	19		5.1			7.4				5.1	3.4	6.3		2.3				2.9	5.1				
Patiala District	7.3		6.3	12	3.7		3.1	4.6	4.4			4.2			3.6						3.2			

Table 1.3: Ten Leading Sites of Cancer City Wise for Females (2012-14)

City	Lung	Mouth	Tongue	NHL	Oesophegus	Gall Bladder	Lymphoid Leuk	Breast	Cervix Uteri	Ovary	Thyroid	Corpus Uteri	Stomach	Brain	Rectum	Vagina	Colon	Uterus Unsp	Hypopharynx	Other Skin	Myeloid Leuk	Nasopharynx	Multiple Myel	Liver	Leuk
Banglore	3.7	3.9			3.5			27.5	12.3	5.3	4	3.7	3.6	2.7											
Barshi Rural	3.6	2.1			3.8			20	16.1	4.6	2.1			1.9	2.1										
Barshi Expanded		1.9	2.8		2			21.8	28.6	4.6			2.7		2.2	2.3	1.8								
Bhopal	3.3	4.9	3.3		3			31.2	12.5	7.8							2.3								
Chennai	3.3	3.1		2.4	2.6			30.7	12.6	6.6			4.1				2.4								
Delhi	3.1	2.2		3.1	2.3		2.1	28.6	10.8	7.2		3.5													
Mumbai	5	3.2		2.8	2.7	3.5		28.8	7.7	6.9		3.3													
Cachar District	3.2	3.9	2.2		6.8	10.3		14.3	13.9	5.1			2.6					2.7							
Dibrugarh District	2.6	4			9.4	10.7		19	6.4	8.9			5.4		2.5				2.7						
Kamrup Urban District	4.4	3.5			10.2	9.3		17.5	8.6	5.5			4.9		2.5		2.6								
Manipur State (MR)	14			3.6		5.2		15.3	9.2	5.7	8		3.3		2.9					3.9					
Imphal West District	14			3.5		5.8		16.5	8.4	6.2	8.9				3.4					4	3.2				
MR-Excl. Imphal West	14			3.7		4.8		14.8	9.5	5.5	7.6		3.6							3.9		3			
Mizoram State (MZ)	16				4.2			13	15.9	2.7	2.4		11.3		2.3		2.3								
Aizwal District	18				4.5	2.6		14.5	15.6	2.9	2.5		10.4												
MZ-Excl.Aizwal	13				3.8			11.3	16.3	2.4	2.2		12.1		2.4							2.2			
Sikkim State	6.8				5.3	6.3		12.7	10	5.9			5.9	4						2.7					
Ahmedabad Urban	3.2	4.8	4.6		4.4	2.3		31.5	9.3	5.3		2.8									2.3				
Aurangabad	4.4	2.8	3.3	2.2	3.3		2.1	30.6	19.9	6.2		2.5													
Kolkata	7.1	3	2.3			7.4		25.4	10.1	7.8	2	3.7	2.5												
Kollam District	3.8	3.9		2.7				27.9	6.8	5.3	11	2.9			3.7								2.4		
Nagpur	2.7	5.1	3.3	1.6	4.1			31.9	13.5	6.3		3.1		2.2											
Pune	4.3	4.4		2.6	3.2			31.4	10.6	7.4		3.8		2.5			2.4								
Thi'puram District	3.9	2.6		2.5				28.5	6	5.9	10	4.3			3		2.5								
Meghalaya	4.2	7.7	2.2		22.2	5		7.9	11.1	2.4			7.4						2.3						
East Khasi Hills District	3.6	7.5	1.9		26.5	5.7		8.4	9.6	2.2			6.8						1.9						
Tripura State	4.6	4.9	2		5.8	9.3		13.7	16.8	6.1			4.2								2.3				
Nagaland	3.1				2.8			12.6	16.7	2.8	5.3		11		2.8							11			
Wardha District	3.2	5.7		2.3	4.6			28.2	13	6.6	4.6									3.4					
Naharlagun (NH)	4.1				3.4	3.6		10.8	12.9	6.8	9.2		13.8							4					
Papumpare District	4.2				4.5	4.5		15	13.5	6.3	11		9.9		2.7										
NH-Excl. Papumpare	4							7	12.4	7.3	7.6									5.1		2.7		8.6	3
Pasighat					3.1	1.9		17.6	23.9	7.6	2.5	2.5	10.7							3.8				3.8	
Patiala District	2.4				5.9	2.8		30.3	10.5	4.9		2.7	2					2.8			2.3				

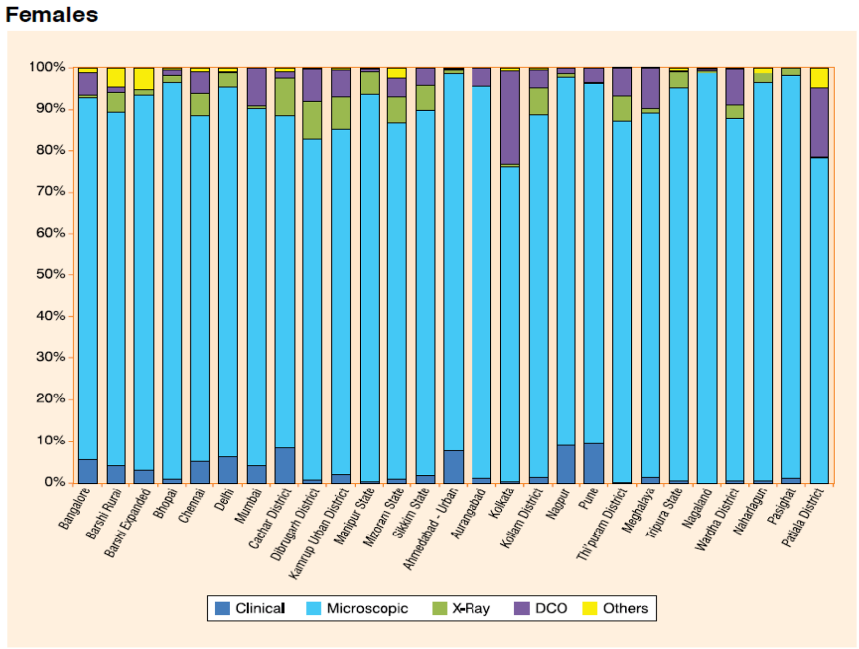


Figure 1.4: Proportion (%) of Cancers based on Different Methods of Diagnosis - All PBCRs for Females

After analysis of data from Figures 1.2, 1.3, and 1.4 we can observe that the most preferable diagnosis type is imaging and microscopic concepts. It is a vibrant area of research. The digital pathology is one of the affective technique which can furnish an efficient and optimized solution.

1.2 Digital Pathology

Cancer is a term used to describe a disorder in human body characterized by abnormal cell proliferation that has the possibilities to infiltrate to other portions of the human body. Oncology is a medical specialty that deals with cancer prevention, diagnosis, and therapy. The following are the three components that have enhanced cancer survival:

i) Prevention - This is accomplished by reducing risk factors such as cigarette and alcohol intake.

ii) Early diagnosis - Common cancer screening, as well as full diagnosis and staging

iii) Treatment - Treatment in a comprehensive cancer center’s and multimodality management by discussion in a tumour board [9]

Due to the technical enhancement and usage of computer science in medical applications, digital pathology has a vital role in this field. Digital Pathology is an image-based vibrant platform that allows pathology information to be acquired, managed, and interpreted from a digitized glass slide. Digital pathology is quickly gaining traction as a proven and necessary technology, with special support for tissue-based research, education, the practice of human pathology, and drug discovery globally. It's a breakthrough dedicated to lowering laboratory costs, increasing operational potential, productivity, and enhancing prognostic decisions and patient care (“Digital Pathology Association” 2013). Life science applications include high production scanning of glass slides, real-time web-based consultations with qualified pathologists, quantitative analysis of entire slide pictures, and secure archiving of pathology data. Figure 1.5, demonstrates the overall working of a traditional and digital pathology.

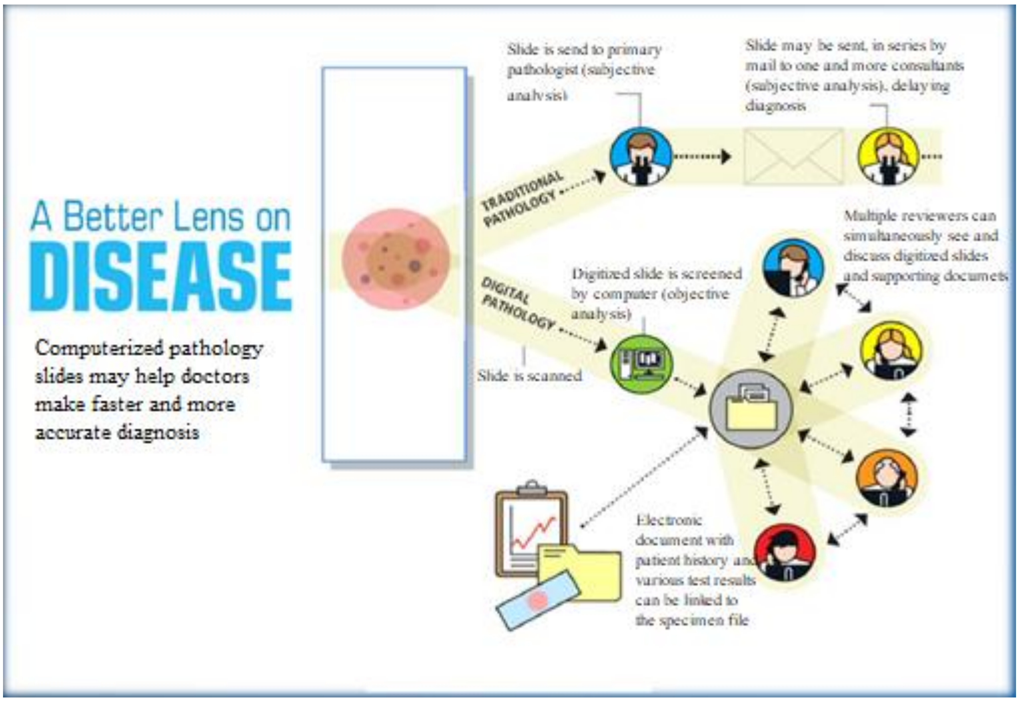

A Better Lens on
DISEASE
Computerized pathology slides may help doctors make faster and more accurate diagnosis
Slide is send to primary pathologist (subjective analysis)
Slide may be sent, in series by mail to one and more consultants (subjective analysis), delaying diagnosis
TRADITIONAL PATHOLOGY
Multiple reviewers can simultaneously see and discuss digitized slides and supporting documets
DIGITAL PATHOLOGY
Digitized slide is screened by computer (objective analysis)
Slide is scanned
Electronic document with patient history and various test results can be linked to the specimen file

Figure 1.5: Digital Pathology Overview

A traditional approach is a non-digitized mode of slides and conventionally it is diagnosed by the experienced pathologist under a microscope. In certain cases, it requires a second opinion that is done by emailing the slides sample to the other pathologists to subjective analysis and delayed the diagnosis results. On the other hand, digital pathology helps in all ways to assist the pathologist to decide within a lesser time duration and multiple reviewers can examine and debate the digitized slides and supporting papers at the same time that can help to make the correct decision [10].

1.2.1 Hardware Support and DICOM Standard

Digital pathology requires high end hardware device like digital scanners of high capacity that enables to scan the sample up to x100 zoom capacity. There are many multinational companies like Leica, Aperio, Hamamatsu, Nikon, Olympus, 3D Histech, Philips, Siemens, etc are the pioneer to manufacture such scanners. These devices are enabled with an array of photosensitive elements (pixels) sensors like CMOS or CCD that convert the optical signals of the slides to digital signals and produce compressed high resolution images following quality standards. To address the challenges of turn-round time and costs in Pathology Operations, it is required customized, integrated, and advanced technological solutions. Integrated Digital slide imaging and management are some of the niche concepts (in WSI Segment), making it easy for hospitals, Research Organizations, Educations Institutes, etc to efficiently and effectively manage cancer diagnostic cases. The solution should enable ease of access for Pathologists to analyze slides with "Clarity & Precision". Whole Slide Imaging (WSI) principles are supported by the "Digital Imaging and Communications in Medicine (DICOM)". Instruments that acquire WSI digital slides should save these images into commercially accessible "Picture Archive and Communication System (PACS)" systems utilizing DICOM-standard messaging to make WSI adoption easier in hospitals and laboratories. After then, the PACS system's competent to store, archive, retrieve, search, and manage these new types of images. Whole slide images are very large. The pathologist cuts the sample that is cancerous affected body part (tissues) for biopsy [11].

1.2.2 Resolution and size of WSI Images

A typical WSI sample can be 20mm × 15mm in size, with a resolution of 0.25 micrometers per pixel (mpp). Most of the optical microscopes have a 10X magnification ocular lenses, utilizing a 40X objective lens results in 400X magnification. Images captured with a resolution of 0.25mpp are referred to as 40X, images captured with a resolution of 0.5mpp are referred to as 20X, and so on, despite the fact that instruments that digitize microscope slides do not use ocular lenses or microscope objective lenses. As a result, the image is around 80000 by 60000 pixels in size, or 4.8 gigabytes. Because most images are recorded in 24-bit color, the image data size is around 15GB. Sometimes the sample sizes up to 50mm x 25mm require higher resolution. In this case, we use z-plane concepts. Data size may vary between 1GB to 1TB depends on cases [12].

1.2.3 Staining and Imaging of Cells

To enhance the visualization of features of the cells or the cellular components under a microscope, it is required the staining of the tissues. After staining it makes it easier to read and visualize the cells by pathologists and they can differentiate the dead and alive as well malignant and benign cells. The different types of staining evolve the different colors for the nuclei of tissues. The various important histological stains are Carmine, Hematin and Hematoxylin, Eosin and Hematoxylin (E&H), Gram Stain, and Trichrome Stains [13]. It can be used various kinds of imaging techniques for a cancer diagnosis like CT Scan, MRI Scan, X-Ray, Mammography, Nuclear medicine scans, and Ultrasound [14].

1.2.4 Cellular and Tissue Level Analysis

Due to the advancements in optical technologies that can generate macro and micro level images. This improved visualization is helping clinicians with the depth observation of the tissues and helping for effective decision making. With the help of various staining techniques and optical technology like WSI concept and the use of fluorescence spectroscopy to detect cervical squamous intraepithelial pre-malignancies and cancers is a promising technology [15]. These technologies helping to raise the patient's standard of living and also reducing the overall medical cost of the diagnosis. The various software tools that have been developed and are supported by artificial intelligence that works at a cellular level are assisting the pathologists and clinicians for improved decisions.

1.3 Computer-Aided Diagnosis (CAD)

CAD was first offered in the early 1990s, barely a few years after the WSI concept was discovered. The CAD system could assist the pathologist in many ways to provide the image measurements that can be used to decide on various diagnoses. The USFDA has now accepted the histological CAD used for diagnosis and also allowed the WSI system for clinical use. It has taken approximately two decades for the WSI system to achieve clinical diagnosis usage [16]. The CAD supports identifying the potential region of interests (ROIs) that helps to reduce the time of diagnosis because the ROIs may cover only a small fraction of the input sample [17]. The other metrics that are supported by the CAD system are cell density [18], mitotic event counts [19], nuclei shape, and other feature metrics that help to identify the malignancy of the nuclei [20][21]. Overall it would be good to say that the CAD system might support early detection and the grade/level of various cancers supported by pathological processes.

1.4 Research Challenges

While the digitization of pathology slides are opening with various advantages but also some challenges discussed below:

1.4.1 Cost and Computational Challenges

Initial infrastructure cost is high for digitizing the pathology slides and its supporting environment like high resolution display, high speed network, good storage capacity for archiving the digital slides, training to the pathologist about the system and scanners, etc. It also requires the high performance computing devices to process and analyze the slides within a short duration. It must be enabled with AI that is a big challenge to localize the ROIs and differentiate between malignant and nonmalignant tissues, finding the features. Because of the high resolution and large number of samples needed for each patient, it takes time to read the samples and execution the supporting models that are again a very challenging factor.

1.4.2 Unavailability of the labeled dataset

This is an interdisciplinary research work and the combination of medical sciences and computer engineering.

1.4.3 The Absence of Gold Standard

For pathology specimens, there is not enough admissible ground truth is another equally difficult issue. Even within the pathology community, the concept of an appropriate and widely accepted gold standard can be contentious. Manual tracing of regions of interest, for example, has been found to be unreliable and should not be used as the gold standard's only source [22]. This demonstrates the importance of getting a multiple opinion that is frequently gained with the help of another person (a subject-matter expert) or computer-aided technology.

Extended patient consequences, in combination with other prognostic methods and biological aspects, it can be examined as an acceptable gold standard. Such kind of processes comprises more costing and time-consuming, and for some circumstances, they might not be available.

1.5 Research Gap and Direction

1.5.1 Problem Definition

The major concern of this research program is to achieve utmost accuracy for finding the “Region of Interest (ROI)” and to improve the performance of existing works.

1.5.2 Research Objective

The objective of the research work is the Development of Pattern Analysis and Machine Learning Technique for Cancer Diagnosis.

Sub Objectives:

i. To study the Whole Slide Imaging technique.

ii. Review of existing works on WSI.

iii. Reading of high dimensional images

iv. Image Viewing/Zooming/Panning and creating digital slides repository.

v. Devising the efficient algorithm for feature extraction and pattern analysis.

vi. Devising algorithm for localizing on Region of Interest. This is the novel unique approach for localizing the ROI based on the frequency of the nuclei

vii. Implementing the newly devised algorithm.

viii. Performance testing of the new algorithm and comparing the existing algorithms to show the efficiency of the new one.

1.6 Research Contribution

1.6.1 Cost-Effective Methodology

The aim of this work is to develop a cost-effective model to identify the region of interest within optimized resources and time duration. The developed methodology will assist the pathologist to do a better diagnosis.

1.6.2 A Novel Approach To Localize Region Of Interest

Fluorescence, immunohistochemistry, and multispectral imaging ideas facilitate whole-slide imaging. Finding ROI on a cancerous sample image is difficult due to the large size of WSI images and its calculation. To discover the relevant region of interest, unsupervised machine learning and computable analysis of cancerous samples are supplemented by morphological characteristics and shape formulas. Due to computational constraints, it is better to start by working on small patches, integrating the data, and automatically detecting or localizing the ROI. It is also compared to the manual and automated ROI of ICIAR2018 dataset.

1.6.3 Evaluation and Prediction of Cancer Grade

It is always a challenging task to analyze automatically the immunohistochemically Ki-67 stained images due to irregular color intensities distributions among different cell types. The other challenge is, immunohistochemical detection of Ki-67 antigen is suffering from a nonstandardized procedure for Ki-67 assessment and interpretation used for the marker's clinical utility. To solve the issues and giving a standardized procedure, unsupervised machine learning techniques like clustering of local associated features are suggested in the proposal. Unlike a traditional approach, the algorithm segments the sample image of the cells based on texture and color space. The segmentation helps to characterize the cells with certain tumor grading de facto criteria for reference to effective and qualitative pathological analysis. The quantitative results of the algorithm showed effective nuclear section segmentation with high accuracy and robustness.

1.7 Thesis Navigation

The thesis is organized into the following chapters. The first chapter brief about the introduction and motivation about the work. It includes the local and global scenario of cancer statistics, digital pathology, Computer Aided

Diagnosis (CAD), research challenges and gaps, research objective, and contribution.

The second chapter brief the background of technologies and literature survey related to breast cancer diagnosis. It includes the topic of application of medical imaging in cancer diagnosis, preprocessing, feature extraction and selection, and implementation of advanced machine learning techniques in cancer diagnosis.

The model for a clinical decision support system is concluded in chapter third. It includes the sub-topics biomarkers, quality control over digital slides, quantitative image description, predictive modeling, visualization, and exploratory analysis.

The fourth chapter is describing a novel approach to localize the ROI in WSI, supported by shape formulas and morphological features on ICIAR 2018 BACH dataset. The fifth chapter is brief about the prognostic evaluation and grading of breast cancer using the BreCaHAD dataset.

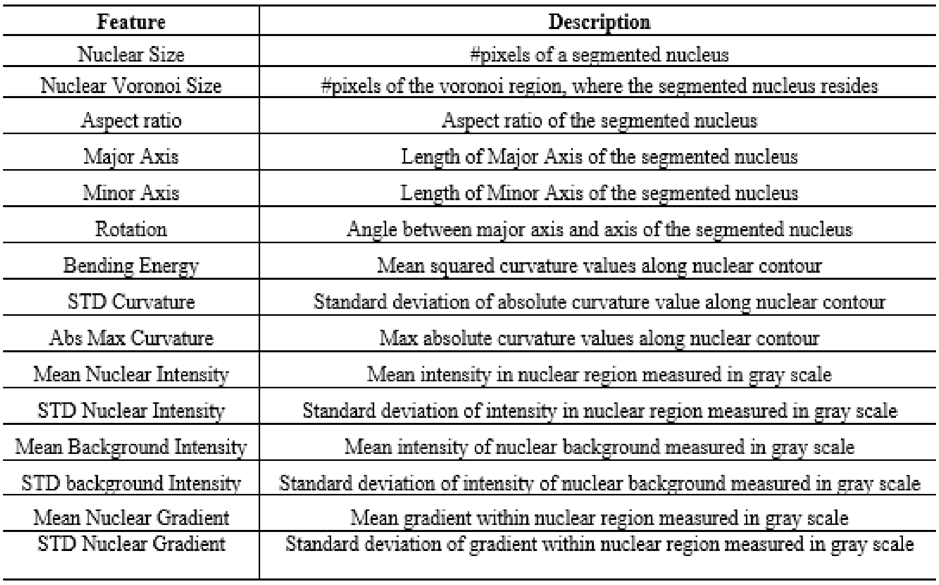
Chapter 2

Background and Literature Survey

2.1 Medical Imaging in Cancer Diagnosis

The data structure used to describe image data can make or break an image processing task's success. The image pyramid is one such structure that has gotten a lot of attention. The image pyramid extends a useful image depiction for many tasks. Pyramid filtering is more efficient to compute than the equivalent filtering done with a quick Fourier transform. Because the nodes at each level reflect information that is localized in both space and spatial frequency, the information is also available in a format that is easy to utilize. [23]. The classification algorithms for tissue histology based on strong depiction of morphometric factors mentioned in Table 2.1, which are developed at nuclear level morphometric features for different positions and scales within the “Spatial Pyramid Matching (SPM)” framework. These methods are implemented and tested on two distinct datasets and tumors gathered from “The Cancer Genome Atlas (TCGA)” [24].

Table 2.1: Morphometric features of tissue and its description



Feature	Description
Nuclear Size	#pixels of a segmented nucleus
Nuclear Voronoi Size	#pixels of the voronoi region, where the segmented nucleus resides
Aspect ratio	Aspect ratio of the segmented nucleus
Major Axis	Length of Major Axis of the segmented nucleus
Minor Axis	Length of Minor Axis of the segmented nucleus
Rotation	Angle between major axis and axis of the segmented nucleus
Bending Energy	Mean squared curvature values along nuclear contour
STD Curvature	Standard deviation of absolute curvature value along nuclear contour
Abs Max Curvature	Max absolute curvature values along nuclear contour
Mean Nuclear Intensity	Mean intensity in nuclear region measured in gray scale
STD Nuclear Intensity	Standard deviation of intensity in nuclear region measured in gray scale
Mean Background Intensity	Mean intensity of nuclear background measured in gray scale
STD background Intensity	Standard deviation of intensity of nuclear background measured in gray scale
Mean Nuclear Gradient	Mean gradient within nuclear region measured in gray scale
STD Nuclear Gradient	Standard deviation of gradient within nuclear region measured in gray scale

2.1.1 Input Images

There are various capturing technology of input images as depicted in Figure 2.1.

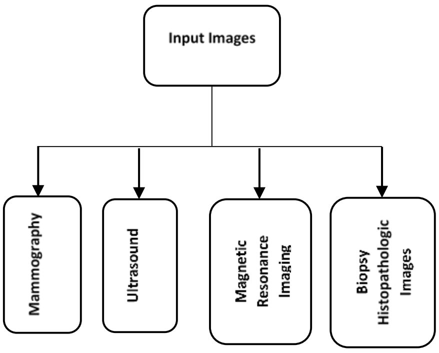


Figure 2.1: Different modes of input images to CAD

2.1.1.1 Mammography

In current scenario, mammography is one of the most efficient method for the identification of breast cancer in its early stages. The sample of mammographic specimens of breast cancer is shown in Figure 2.2. It has some edges and challenges. Casualty reduction, enhanced early illness treatment, enhanced quality assurance of the diagnostic chain are the effectiveness of mammography. As a result of radiation dangers, the risk of false alarm or a FP and FP alarm, radiologists miss 10% to 30% of breast cancinoma and interval cancinoma, and overdiagnosis are the major demerits of mammography [25].

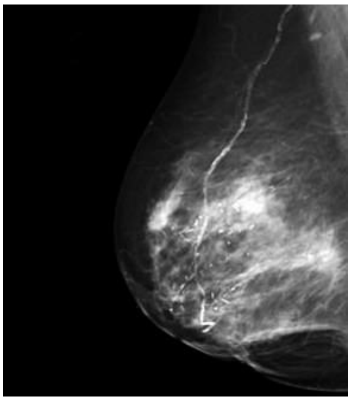


Figure 2.2: Representing the mammographic images of breast cancer

Tomosynthesis, are observed at various angles that help to produce thin cross-sections. It is developed as 3D mammography that has been consented by “Food & Drug Administration (FDA)” to screen the breast carcinoma in which x-rays of the 3D breast images using standard CT scan technology. It is applying various state-of-the-art ML models to effectively identify the breast carcinoma to exert mammograms. Deep Learning models are applying on specimens from the INbreast data repository where initial training needed lesion annotations that has an AUC of 0.95 per image, with four-model averaging increasing the accuracy to 0.98 [26]. Another cutting-edge breast lesion identification technique is based upon a non-ionizing technology that uses dielectric characteristics to distinguish between cancerous and non-cancerous tissues. A microwave equipment used to gather clinical data of breast lesion specimens. These specimens are used to train the various classifiers likes “KNN (k-nearest neighbor)” algorithm, “Multi-Layer Perceptron (MLP) Neural Network”, and “SVM (Support Vector Machine)” achieved the prediction accuracy of 98% [27].

2.1.1.2 Ultrasound

Ultrasound is a systematic technique to assess the breast carcinoma and it can be suggested during lactation and pregnancy. In case of dense and smaller breast ultrasound is more suggested technique to compare various mammography results. It is always preferred before biopsy to reduce inessential biopsy. The color Doppler Imaging and ultrasound echo-enhancing yield further knowledge that helps to differentiate cancerous and non-cancerous samples with good accuracy [28]. Such type of CAD system is very useful for interpreting images and training of junior radiologists.

The radio frequency signals gathered from the tumor and its environments and the information fetched are compile to find out the quantitative measures like texture parameters, shape parameters, and entropy. A multi-parametric classifier are implemented and attained an AUC of 0.83 and 0.92 for inner and outer tumor data respectively [29]. The sample of normal, benign, and malignant ultrasound images are shown in Figure 2.3.

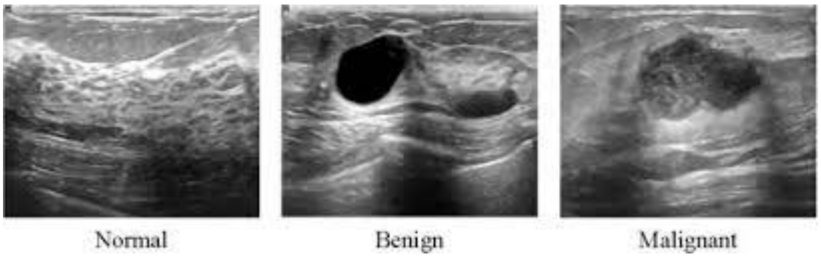


Figure 2.3: Representing the ultrasound image samples of different labels like Normal, Benign, and Malignant

2.1.1.3 Magnetic Resonance Imaging (MRI)

Breast MRI technique is an eventual substitute, however, due to the expensive cost, it is not generally available as mammography and ultrasound. MRI is advised for screening high-risk women for breast cancer, as well as analyzing suspicious areas identified by mammography to assess the size of the lump as illustrated in Figure 2.4. MRI acquisition does not use ionizing radiation. The process of interpreting MRI sample images takes a long duration and need a lot of radiologist knowledge to discover and distinguish malignant and benign tumors [30]. MRI and evolution in a 3D printed surgical guide produced by a 3D printer marking the primary tumour is really useful. MRI has yielded as one of the accurate method to assess the residual tumors after neoadjuvant chemotherapy [31].

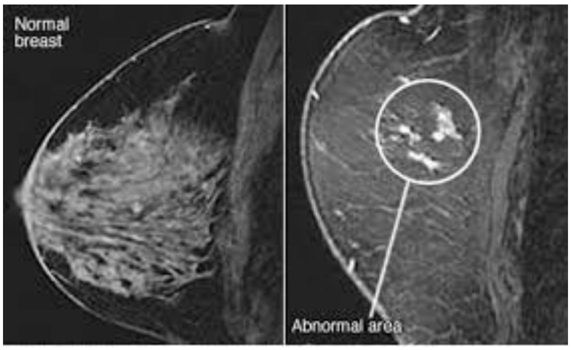


Figure 2.4: Representing the MRI images of breast cancer

2.1.1.4 Biopsy Histopathologic Images

When mammography or other modalities reveal any form of abnormalities, the biopsy is considered the final step. “Fine-Needle Aspiration Biopsy (FNAB)”, surgical biopsy, and core biopsy are all types of biopsies in which a specimen is taken from a suspected lesion and inspected under the microscope by a skilled pathologist. Revolutionaries in digital pathology the field of computer vision-based biopsies, in which physical slides are transformed into digital slides. It can also be defined as WSI that are scanned by high competent scanners in Figure 2.5.

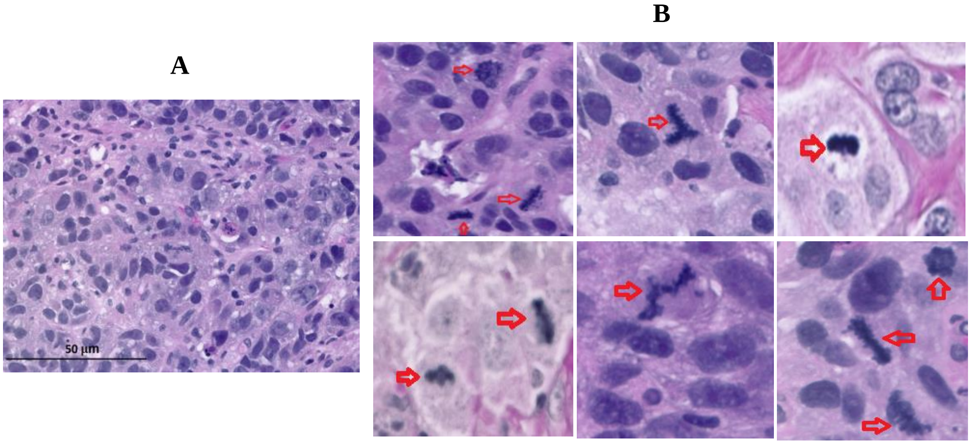


Figure 2.5: Representing the H&E histopathologic image sample in A. The red arrow showing the mitosis figures, rest are tumor nuclear sections in B

. It is possible to obtain numerous opinions for aberrant samples within hours using a virtual microscopic concept. For the same issue, obtaining a second opinion requires a physical sample that can take several weeks or months. Immunostaining procedures such as IHC, H&E, and others help to color the images so that they are more readable by pathologists and the system. Hematoxylin stains turn the nuclei of the cells blue, while Eosin stains turn the tissue pink. There are a variety of deep learning algorithms that may be used on different datasets to categorize them with an accuracy of more than 90% [32]. Pathologists believe that H&E will continue to be one of the best practices for the next 50 years, based on a lengthy history of H&E, a vast amount of data produced, multiple state-of-the-art procedures, and high-quality research publications [33]. The CAD system in medical imaging is developed to detect, diagnose, and forecast disease prognosis using digitized histopathological specimens. The lack of a clinically annotated histopathology dataset to train the various classifiers is a significant impediment in achieving accurate and effective findings. For processing and training high dimensional histopathology or WSI images, high computing systems such as graphics processing units (GPUs) are necessary [34].

2.1.2 Image Pre-processing

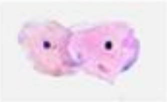
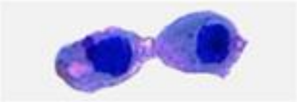
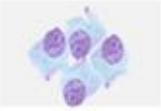
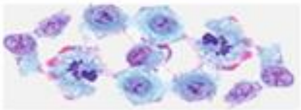
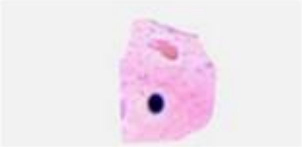
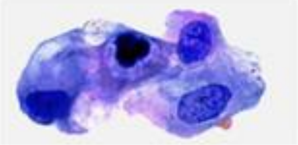
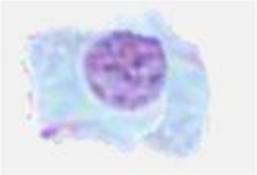
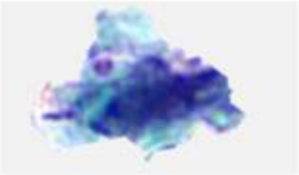
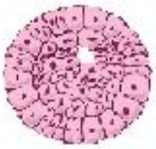
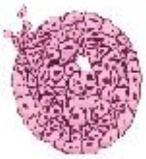
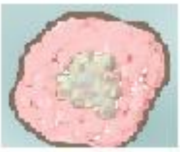
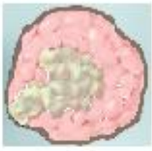
The role of image pre-processing is critical in CAD to produce optimal results. Through the frequency and spatial domains, this phase focuses on image resizing, noise removal, and image intensity improvement. Color and illumination normalization is one of the first and most significant techniques for both fluorescent and bright field histopathology images. Normalization aids to lower the dissimilarity in tissue specimen due to the disparity in staining and scanning surroundings. By applying the illumination pattern or using calibration targets, fitting polynomial surfaces can be used to rectify illumination disparities [35]. Other methods for correcting spectral and spatial illumination differences include histogram equalization and matching. Various color spaces, such as HSI, LUV, and CMYK, can also be utilized to improve an image's quality during preprocessing.

2.2 Feature Extraction and Selection

There are so many algorithms for feature extraction. A “Dominant Local Binary Patterns for Texture Classification” that an innovative method for extracting picture characteristics for texture classification. The directed attributes are less susceptible to histogram equalization and noise, and are robust to image rotation. It is made up of two sets of features: DLBP in a texture image and additional attributes take out from circularly symmetric Gabor filter outputs. The presiding local binary pattern method captures descriptive textual information by using the most often occurring patterns, whereas Gabor-based features try to supplement the DLBP features with extra global textural information [36].

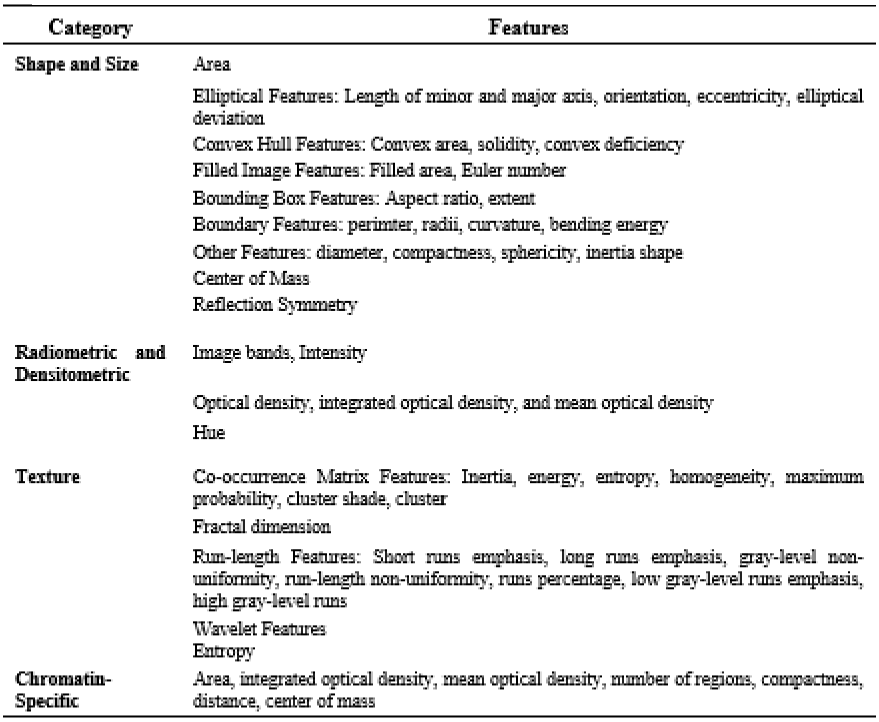
It is listed in Table 2.2, the distinction between malignant and non-cancerous cells. Mostly based on features, color, and structure differences, it can be formulated the differences. Image traits are taken from the ROI after segmentation to detect and assess possible malignancies. One of the most crucial processes in the examination of biopsy pictures is feature extraction. For better predictions, characteristics are retrieved at the tissue and cell levels of microscopic biopsy pictures. It extract the shape characteristics like anticircularity, contour irregularity, and area irregularity of nuclei to throw back the asymmetry of nuclei in biopsy specimen using both contour and region-based approaches to capture the shape data more effectively.

Table 2.2: Difference between Normal and Cancerous cells

Normal Cells	Cancerous Cells	Description of cancerous cells
		Large and variably shaped nuclei
		Many dividing cells and disorganized arrangements
		Variation in size and shape of nuclei
		Loss of normal feature (shape and morphology)
		Non invasive cells and invasive cells
		Left is the immunonegative and right one is the immunopositive sample of cells

The selected feature target to assess the behavior of each cell at the cellular level, without taking into consideration spatial dominion between them. The shape and morphology, textural, histogram of wavelet features, and directional gradients are retrieved from biopsy images for each cell. The tissue-level characteristics aid in quantifying the cell distribution throughout the tissue, and they generally rely on either the grey level reliance of the pixels or the spatial dependency of the cells [37].

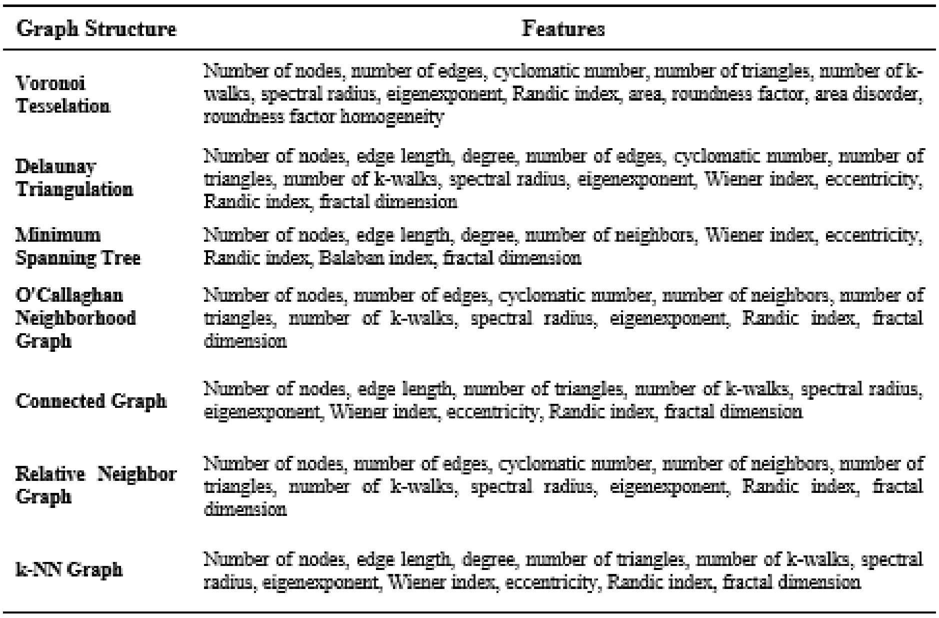
Table 2.3: Summary of the object-level characteristics used in histopathological image analysis



Category	Features
Shape and Size	Area
	Elliptical Features: Length of minor and major axis, orientation, eccentricity, elliptical deviation
	Convex Hull Features: Convex area, solidity, convex deficiency
	Filled Image Features: Filled area, Euler number
	Bounding Box Features: Aspect ratio, extent
	Boundary Features: perimter, radii, curvature, bending energy
	Other Features: diameter, compactness, sphericity, inertia shape
	Center of Mass
	Reflection Symmetry
Radiometric and Densitometric	Image bands, Intensity
	Optical density, integrated optical density, and mean optical density
	Hue
Texture	Co-occurrence Matrix Features: Inertia, energy, entropy, homogeneity, maximum probability, cluster shade, cluster
	Fractal dimension
	Run-length Features: Short runs emphasis, long runs emphasis, gray-level non-uniformity, run-length non-uniformity, runs percentage, low gray-level runs emphasis, high gray-level runs
	Wavelet Features
	Entropy
Chromatin-Specific	Area, integrated optical density, mean optical density, number of regions, compactness, distance, center of mass

Shape, texture, and color-based descriptors are some of the picture descriptors utilized for feature extraction. To locate the required features, various filters, fractals, and morphological operators can be utilized. It is always necessary to choose the suitable one so that a classifier can predict the label of the input sample image with ease. The memory size, computational cost, and robustness of the effective feature extraction technique should always be optimized. Gabor wavelet-based algorithms can help diagnose breast cancer by extracting aberrant features from mammography specimens [38]. Most of the time the important features for breast carcinoma detection using histopathological specimens are defined by clinicians. The development of computer vision methods is critical that will be efficient of object level and spatial-relation characteristics analysis based on graph structure as depicted in Tables 2.3 and 2.4 [34].

Table 2.4: Summary of spatial order of characteristics used in histopathological image observation



Graph Structure	Features
Voronoi Tesselation	Number of nodes, number of edges, cyclomatic number, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, area, roundness factor, area disorder, roundness factor homogeneity
Delaunay Triangulation	Number of nodes, edge length, degree, number of edges, cyclomatic number, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension
Minimum Spanning Tree	Number of nodes, edge length, degree, number of neighbors, Wiener index, eccentricity, Randic index, Balaban index, fractal dimension
O'Callaghan Neighborhood Graph	Number of nodes, number of edges, cyclomatic number, number of neighbors, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, fractal dimension
Connected Graph	Number of nodes, edge length, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension
Relative Neighbor Graph	Number of nodes, number of edges, cyclomatic number, number of neighbors, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, fractal dimension
k-NN Graph	Number of nodes, edge length, degree, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension

2.3 Implementation of Machine Learning Algorithms in Cancer Diagnosis

Machine learning started to use for various applications in the early 1990s. Arthur Samuel nicely defined it as giving “computers the ability to learn without being explicitly programmed”. Nowadays machine learning algorithms have a wide scope in healthcare especially in CAD for automatic diagnosis of various major diseases including cancer. Machine learning algorithms are used in CAD to distinguished and classify abnormalities. It is the most important component of CAD that helps to an automatic diagnosis of various cancer diseases.

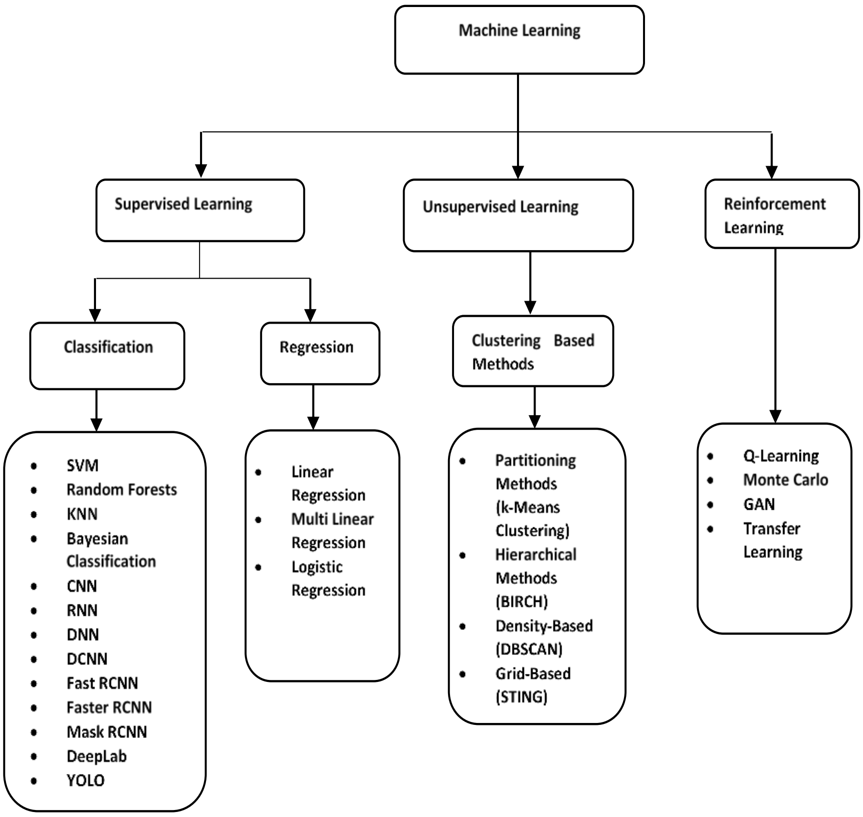


Figure 2.6: Hierarchy of Machine Learning Algorithms

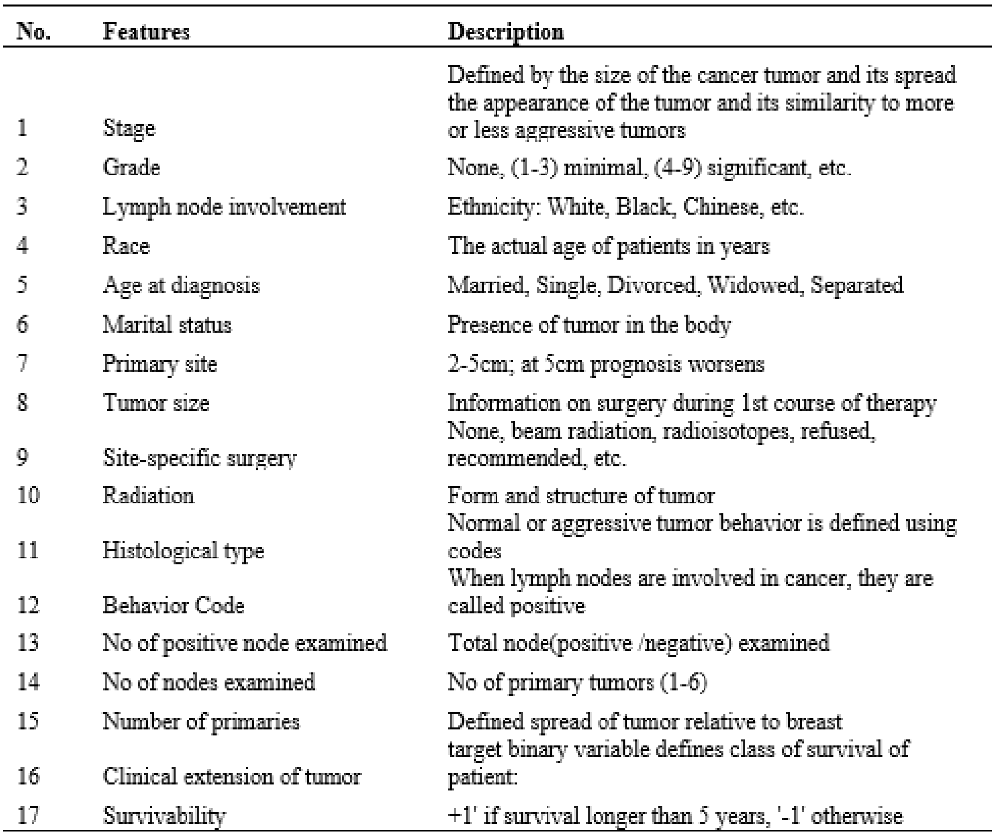
Machine learning is primarily have three classes: supervised learning, unsupervised learning, and reinforcement learning. Many state-of-the-art ML models for classification are achieving accuracy levels of over 90% shown in Figure 2.6 [39].

2.3.1 Supervised Learning Algorithms

Supervised learning is a data processing task in which the training data is annotated and labelled appropriately. It is always necessary to have a sufficient amount of training data that covers all of the labels in a balanced manner, as well as data for testing and validation. The user supplied ground truth make up the training dataset. If the training data is insufficient or unbalanced, the output will be poor, and the performance will suffer as a result. The problem of overfitting and underfitting could be a concern. Regression and classification are two highly related prediction methods under supervised learning as shown in Figure 2.6. The regression models are suitable for continuous target datasets and classification models used for a finite set of values [40]. It is defined the parameters and using a predictive model for evaluating breast cancer survivability. The goal of this project is to discuss the importance of stability and to propose a viable paradigm. The SVM, RF, and ANN Machine learning algorithms have been deployed and compared based on performance. In this study, it is defined the prognostics elements for breast cancer survivability shown in Table 2.5 [41].

Most of the elements are related to behavior and features and based on this the final grade of cancer and the survivability period will be predicted. For convolution neural network training, an expectation-maximization (EM) based algorithm is suggested that automatically selects discriminative patches.

Table 2.5: Prognostics elements for breast cancer survivability



No.	Features	Description
1	Stage	Defined by the size of the cancer tumor and its spread the appearance of the tumor and its similarity to more or less aggressive tumors
2	Grade	None, (1-3) minimal, (4-9) significant, etc.
3	Lymph node involvement	Ethnicity: White, Black, Chinese, etc.
4	Race	The actual age of patients in years
5	Age at diagnosis	Married, Single, Divorced, Widowed, Separated
6	Marital status	Presence of tumor in the body
7	Primary site	2-5cm; at 5cm prognosis worsens
8	Tumor size	Information on surgery during 1st course of therapy
9	Site-specific surgery	None, beam radiation, radioisotopes, refused, recommended, etc.
10	Radiation	Form and structure of tumor
11	Histological type	Normal or aggressive tumor behavior is defined using codes
12	Behavior Code	When lymph nodes are involved in cancer, they are called positive
13	No of positive node examined	Total node(positive /negative) examined
14	No of nodes examined	No of primary tumors (1-6)
15	Number of primaries	Defined spread of tumor relative to breast
16	Clinical extension of tumor	target binary variable defines class of survival of patient:
17	Survivability	+1' if survival longer than 5 years, '-1' otherwise

2.3.2 Unsupervised Learning Algorithms

Unlike supervised learning, where the training dataset is labelled, unsupervised learning does not have a labelled training dataset. Clustering algorithms are the most prevalent type of unsupervised learning. Clustering is the division of a collection of data objects into bunch of clusters, with objects within each cluster having a great deal of similarities while being substantially different from objects in other clusters. There is a wide application of unsupervised learning and healthcare is one of them. Clustering is very useful to find the patterns and outliers. The k-means clustering is a kind of partitioned based algorithm and one of the most famous clustering techniques. It employs a heuristic approach to swiftly arrive at a local optimum [40]. The k-means clustering based segmentation is used for breast cancer diagnosis by distinguishing malignant from benign cases. The RGB color space is supported by adaptive thresholding that was helping to separate the red blood cells from other objects with good accuracy [42]. One of the most crucial processes in determining the ROI is segmentation. Separating one region to another region is based on either color space or texture. The color space LAB has been used and applied color clusters. All color attributes are allowed and employed for segmentation, and it recognizes the respective cross correlation intrinsically [43]. The scope of unsupervised learning is limited and it is primarily used for segmentation based on color and other features.

2.3.3 Classical Machine and Deep Learning Algorithms

Classical machine learning is successfully adopted by the various expert system and AI-based tools used in medical science for years. The importance of expert input and evaluation in these algorithms cannot be overstated. Classical machine learning can be used for anomaly detection, correlation, finding the ROI, disease diagnosis, cancer grading, and prediction, etc. These algorithms use the subject knowledge and the large training dataset to explore and find the strong patterns and levitate them to achieve the prediction tasks.

As previously said, one of the key drawbacks of trivial machine learning algorithms, it requires domain experts. Otherwise, the labeling of the training data will be not authentic and the model can wrongly predict. Because of this constraint, its forecasting capacity is saturated, making it only suitable for specific datasets and inefficient for new datasets derived from unlike tissue types or in a dissimilar form, system, or environment.

In the last decade, there is a great evolution in machine learning algorithms and now it became more data-driven and dynamic in behavior. This evolution is also supported by high-performance computing. Deep learning is a subtype and advance version of machine learning, it can automatically locate the representation required for identification or classification that allowing the model to straight map and input an image sample to an output vector.

A convolution neural network is implemented on whole slide images. The automatic recognition of cancer subtypes, after training a CNN to Whole Slide Tissues Images (WSI) of gigapixel resolution is computationally impossible in current hardware limitations [44]. It is proposed to discuss and underline the impact of a large public database of histopathological data like TCGA. It is implemented the de novo solutions which are based on feature extraction that apprehend features of an image at pixels, semantic, and object levels. It utilizes the image properties for diagnostic or prognostic purposes. It is a kind of clinical decision support system that enables image processing capabilities in place of clinical decision-making thorough data analysis [45].

Digital pathology and histologic image analysis are supporting to transform the practice of pathology increasingly towards a quantitative science. Both the concept should be co-evolve and develop a mutually beneficial connection that yields high-quality, repeatable objective data. Image analysis should be treated as a complementary of trivial histopathology evaluations. The combination provides a complete understanding of the pathologic processes and experimental steps with the support of morphologic changes [46]. Another paper related to deep learning-based algorithms investigates the concepts through different unique digital pathology factors like nuclei segmentation, mitotic detection and counts, tissue classification either cancerous or noncancerous to produce similar in many cases and better outputs from the various state-of-the-art manually feature-based classification algorithms. The calculated F-score of nuclei segmentation is 0.83 and for mitosis detection, it is 0.53 [47].

Various computer-assisted diagnosis (CAD) algorithms support pathologists to decide within less throughput time. Mammography, MRI, Ultrasound, and biopsy histopathologic images are the different modes of input images to CAD. When mammography or other modalities reveal any form of abnormalities, the biopsy is considered the final step. Biopsies, such as "Fine-Needle Aspiration Biopsy (FNAB)", "core biopsy," and "surgical biopsy," are procedures in which a sample is taken from a suspicious lesion and examined under the microscope by a pathologist. Revolutionaries in digital pathology the field of computer vision-based biopsies, in which physical slides are transformed into digital slides. It's also known as WSI images scanned using high-capability scanners as shown in Figure 2.7. It is possible to obtain numerous opinions for aberrant samples within hours using a virtual microscopic concept. The physical sample takes a few weeks, if not months, for a second opinion on the same issue. IHC and H&E are used to colorize images so that they are more understandable by pathologists and the system. Hematoxylin stains turn the nuclei of the cells blue, while Eosin stains turn the tissue pink. There are a variety of deep learning algorithms that may be used on different datasets to categorize them with an accuracy of over 90%[32]. The pathologists believe that H&E will continue to be one of the best practices for the next 50 years, based on a lengthy history of H&E, a vast amount of data generated, multiple state-of-the-art methods, and high-quality research publications [33]. Using digitized histopathology pictures, a CAD system in medical imaging been developed for disease identification, diagnosis, and prognosis.

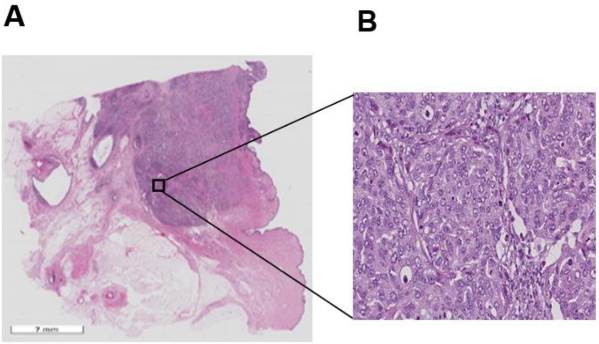


Figure 2.7: A) Specimen of WSI image, B) Zoomed the ROI by x40

The lack of a clinically annotated histopathology dataset to train the various classifiers is a significant impediment in achieving accurate and effective findings. For processing and training high dimensional histopathology or entire slide images, high computing systems such as graphics processing units (GPUs) are necessary [34]. Another study was performed using computer-aided image analysis (CAI) on 1150 H&E stained images collected over 230 different patients suffering from “Invasive Ductal Carcinoma (IDC)” of the breast. It is used the pixel-wise SVM and the pathologist successfully extracted the prognostic information from H&E stained image samples [48]. “The Tumor Proliferation Assessment Challenge (TUPAC)” was conducted in 2016, one of the top three teams were used to build a CNN model to automatically detect the mitotic patterns in H&E stained breast cancer tissue using whole slide images [49]. BACH challenge was organized in 2018 to uplift the state-of-the-art methods for automatic detection and classification of breast cancer tumors using 400 H&E stained microscopy images and 30 H&E whole slide images. The winner of this competition used the CNN model to obtain an overall accuracy of 87%, which was better than several other state-of-the-art algorithms [50].

2.4 Research Challenges in CAD

CAD can assist pathologists and clinicians for the fact, accurate, efficient, and cost-effective decision making diagnosis of the diseases. Presently, CAD is an essential tool for various disease diagnoses especially for cancer after the evolution of digital histopathologic images. The CAD system still has some issues to deal with and that can be rectified by the researchers in the future and it is a continuous process. The list of challenges is [51]:

i) Data gathering and quality control
ii) Development of advanced segmentation algorithms for medical imaging
iii) Development of advanced feature extraction and selection algorithms
iv) Development of better classification algorithms
v) Dealing with big data that include high dimensional images
vi) Development of standard performance assessment technique for CAD system
vii) Lack of gold standards for clinical practice and data usage

2.5 Chapter Summary

This chapter, it is explained the background and literature survey of the technologies supporting the CAD for different cancer diseases. It is explored the applications of digital medical images like ultrasound, mammography, MRI, and biopsy histopathologic specimens. It is discussed the application of various state-of-the-art machine learning and deep learning algorithms and their implementation in cancer diagnosis. The objective and research gaps are listed in the chapter.

Chapter 3

Model for Clinical Decision Support System

The major concern of this research work is to achieve utmost accuracy for finding Region of Interests (ROIs) and improve the performance of existing works. It is a very challenging task to visualize the very high-dimensional images on a screen. Archiving and retrieving such images is again a bigger challenge. It requires good hardware support. Initially, the focus will be reading the WSI images and implementing the various functions like zooming, panning, and sample view with target pointer. The clinical decision support system (CDSS) is an integrated part of CAD and has a very important role to assist pathologists or clinician's decisions.

While diagnosis the cancer patients, pathologists do the observation of tumors and follow the biopsy-derived tissue slides. Based on their skill set and experience, it is tried manually to identify the most affected patches and inspect nuclear morphology, cellular properties, etc. It is eye crying and tedious task to manually examine the tissue samples containing millions of cells with different morphological behavior. It is also subjective and time consuming. To overcome such issues and to support or assist the pathologists or clinicians, it is observed to develop an efficient clinical decision support system (CDSS) as shown in Figure 3.1. The importance and demand of such systems led to several commercial CAD tools for the analysis and diagnosis of cancer disease. The list of the tools and the respective companies which has the copyright are GENIE from Aperio, AQUA Analysis of HistoRx, HALO of Indica Labs, and Visiopharm of Hoersholm. These all tools provide basic image processing capabilities but now due to the evolution in machine learning especially deep learning models, the tools are adopting advanced models to improve the prediction and decision making, also finding the region of interest. None of the above instrument come up with the absolute data analysis for clinical conclusion purposes. It is always required an authentic and large database for such a clinical decision support system. It is led by “NCI Cooperative Prostate Cancer Tissue Resource” [52], the “Human Protein Atlas” [53], and the “NIH Cancer Genome Atlas” (TGCA) [54] for the establishment of a global cancer database.

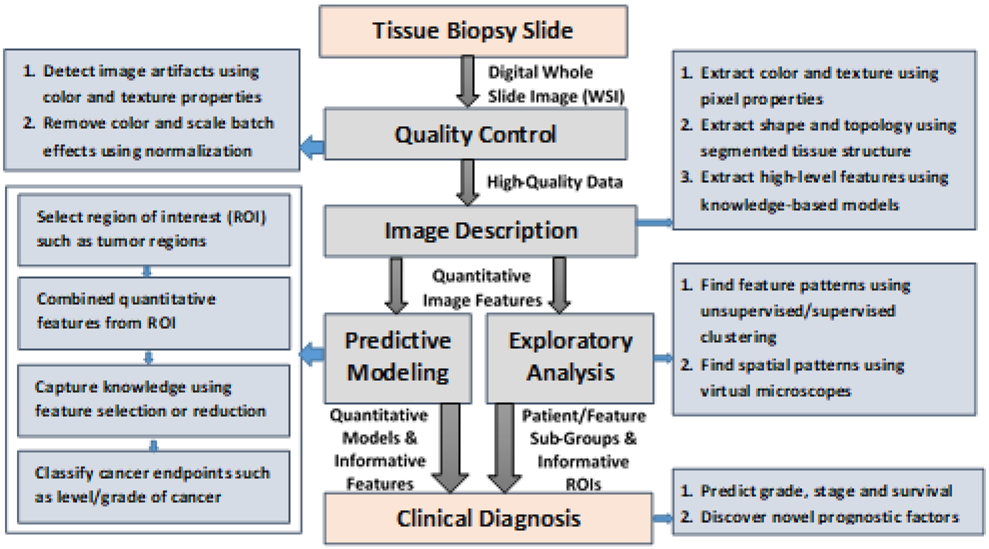

Tissue Biopsy Slide
Digital Whole Slide Image (WSI)
Quality Control
High-Quality Data
Image Description
Quantitative Image Features
Predictive Modeling
Exploratory Analysis
Quantitative Models & Informative Features
Patient/Feature Sub-Groups & Informative ROIs
Clinical Diagnosis
1. Detect image artifacts using color and texture properties
2. Remove color and scale batch effects using normalization
Select region of interest (ROI) such as tumor regions
Combined quantitative features from ROI
Capture knowledge using feature selection or reduction
Classify cancer endpoints such as level/grade of cancer
1. Extract color and texture using pixel properties
2. Extract shape and topology using segmented tissue structure
3. Extract high-level features using knowledge-based models
1. Find feature patterns using unsupervised/supervised clustering
2. Find spatial patterns using virtual microscopes
1. Predict grade, stage and survival
2. Discover novel prognostic factors

Figure 3.1: Model of CDSS for computable analysis of WSI samples of tissue biopsy specimen

3.1 Digital Imaging

WSI is a type of Digital Imaging and is becoming more relevant for clinical trials, especially for cancerous tissues. It has the perspective to be employed in teleconsultation, tele pathology, digital pathology, clinical education, quality assurance, training, and digital image analysis to assist clinicians and pathologists [55]. Virtual Microscopy (VM) supports this concept. VM is a synthesis of optical microscope images, transmitting over a computer network, archived on disks. This concept is very helpful for second opinions or consultations and also solves the problem of the storage of physical tissues through archiving. The focus of the solution is image acquisition, store locally or remotely, transmission on the web, sharing images, viewing images, image analysis, reporting, and archiving. Virtual microscopy has a high potential for various activities in training, pathology education, quality control, and clinical meeting activity [56].

3.1.1 Biomarkers

Biomarkers are becoming increasingly relevant in the clinical care of cancer patients as genetic profiling technology and molecular medicines become more widely available. A biomarker is a biological metric that can be used to describe an organism's normal or pathological biological state. Its presence in the body has the potential to impact or forecast the disease's occurrence **[57].**

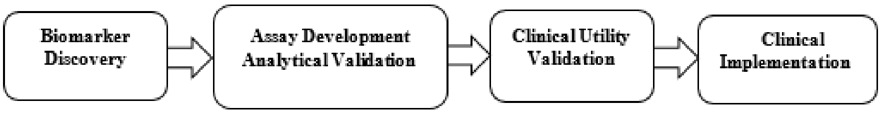

Biomarker Discovery
Assay Development Analytical Validation
Clinical Utility Validation
Clinical Implementation

Figure 3.2: The process of cancer biomarker growth

From early discovery to validation and clinical adoption, biomarker development is a multi-step process as shown in Figure 3.2 [58]. Breast cancer biomarkers are divided into various categories. The predictive biomarkers that predict the responses to specific therapeutic interventions like HER2, KRAS. The prognostic biomarker helps physicians regarding the risk of clinical outcomes such as disease progression in the future or rate of impact. The conventional biomarkers generally suggested in all breast cancer patients are HER2, ER-alpha, , Ki-67, PR, and Histological grade [59].

3.2 Quality Control over Digital Slides

During image acquisition, various artifacts and batch effects can occur as a demerit that can influence the quality of histopathology WSI specimens. The whole digitized image file generally occupies in the range of 1GB to 20 GB of storage space. The file format of the sample images may be either TIFF or JPEG2000, a compressed one sand follows the standards of DICOM.

3.2.1 Image Artifacts

During the preparation of biopsy slides, there may be possibilities of anomalies led by mishandling of microscope or due to the wrong adjustment of its parameters, which is called image artifacts in WSI. The most common image artifacts are air bubbles, pen marks, shadows, blurred regions, mounting media with dirt, tissue folds, and the edge of coverslip as shown in Figure 3.4. If any such kind of artifact arises, the slide will be unsuitable for assessment and required rescanning [60]. These image artifacts can be eliminated by using different filters derived in image processing and unsupervised learning. The image artifacts are also represented as out-of-focus (OOF). ConvFocus based upon CNN model has been created to comprehensively localize and assess the importance of OOF sections on WSI slides. The patch level AUC achieved by the ConvFocus model is 0.95 [61].

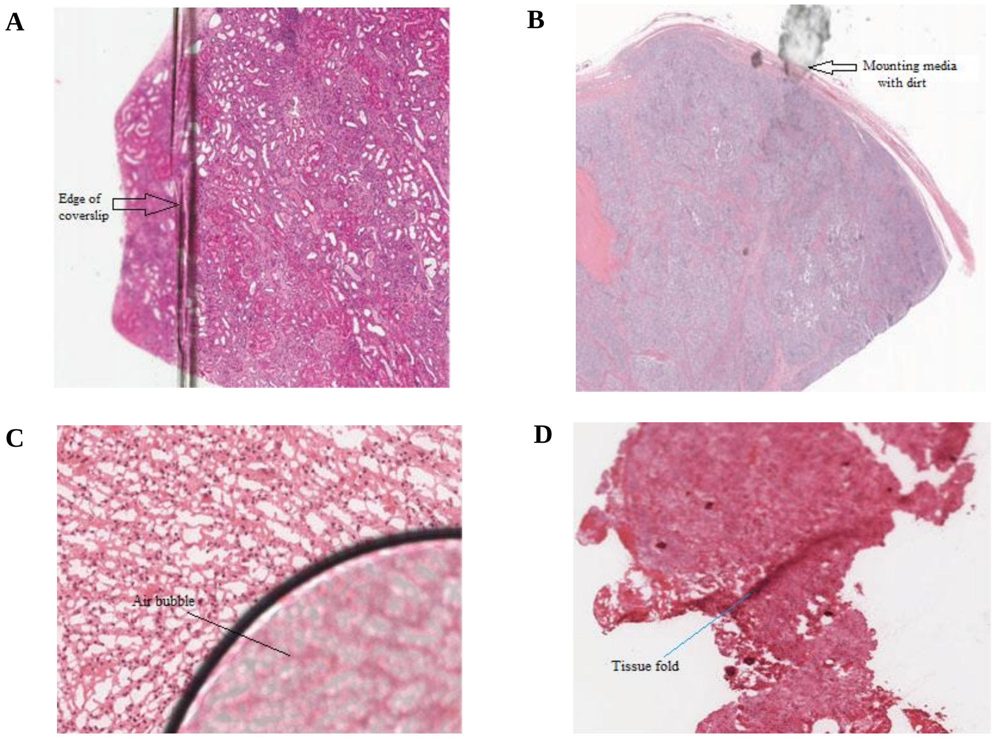


Figure 3.3: Showing the different types of image artifacts (A) edge of overlap, (B) Mounting media with dirt, (C) Air bubble, and (D) Tissue fold

3.2.2 Batch Effects

Variances in picture attributes between two batches could be due to dissimilarities in slide preparation, microscopy, and digitizing device. These discrepancies, known as batch effects, might cause predictive model performance estimates to be skewed. Color and scale batch effects are common in histopathological pictures. Color batch effects can be eliminated by normalizing an image's color at the pixel level. The other batch effects are object size, topology, and texture [62]. The changes in the distribution of image features between batches can be used to detect batch effects.

3.3 Quantitative Image Description

Content-based picture attributes are also carried in WSI data. For quantitative prediction modelling and exploratory analysis, the content-based features are useful. Pixel, object, and semantic features are the three layers of classification [63].

3.3.1 Pixel-level Features

Pixel is the smallest unit of a digital image. Pixel-level image features are at the bottom of the information structure. All image pixels carry the color and textures. The different color spaces are RGB, HSV, CMY, CIELUV, and CIELAB. The texture features measure image contrast, intensity change, sharpness, and edge discontinuities using sudden changes in grayscale values. The other feature selection algorithms like GLCM, wavelets, fractals, and Gabor filters can be used [64].

3.3.2 Object-level Features

Object-level features provide more information than pixel-level features. It discusses the cellular structure like nuclei, glands, and cytoplasm in a WSI sample. To extract the information, it is required to segment the cellular structure of the object. The segmentation can be done based on color and texture. It may be automatic and semi-automatic. To improve the precision of segmentation, it is used pixel neighborhood properties like object graph, graph cut, Markov model, etc. Object-level features associated with shape, spatial and texture information in a WSI [62]. Shape-based characteristics can widely classified into the region and contour-based features. Region growing algorithm is used for segmentation and it is based on contours, shape number, perimeter, boundary fractal, etc. [65]. The features of object-level texture are comparable to those of pixel-level texture, but it captures only the associated subset image pixels of the tissue object. Nuclear features can help in cancer grading, subtyping, and separating malignant cells [64].

3.3.3 Semantic-level Features

A semantic-level feature is normally a statistical order based classification on a group of low-level characteristics like color, nuclear texture, and gray-level distribution. Generally, it requires preprocessing to catch the semantic feature. For semantic characteristics, the BoF (bag-of-features) algorithm is most widely employed. It always required a good number of annotated training data. Consequently, there is a scarcity of research on semantic-level descriptors for histopathology [66].

3.4 Predictive Modeling

Predictive modeling is one of the most important parts of CDSS. The WSI prediction modeling has various phases: ROI selection, selection and reduction of features, as well as classification.

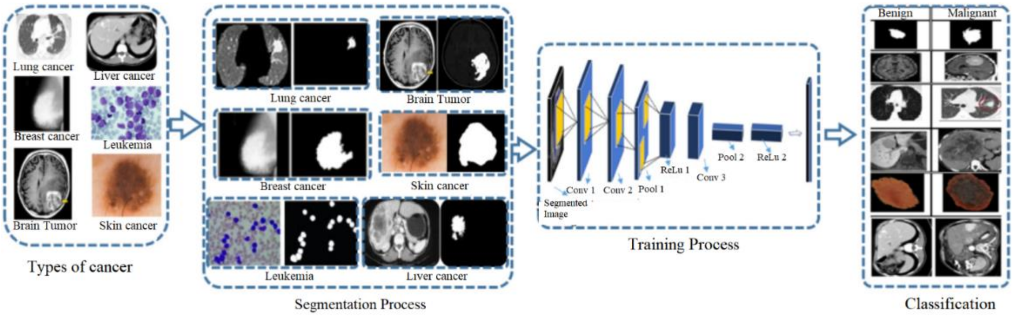
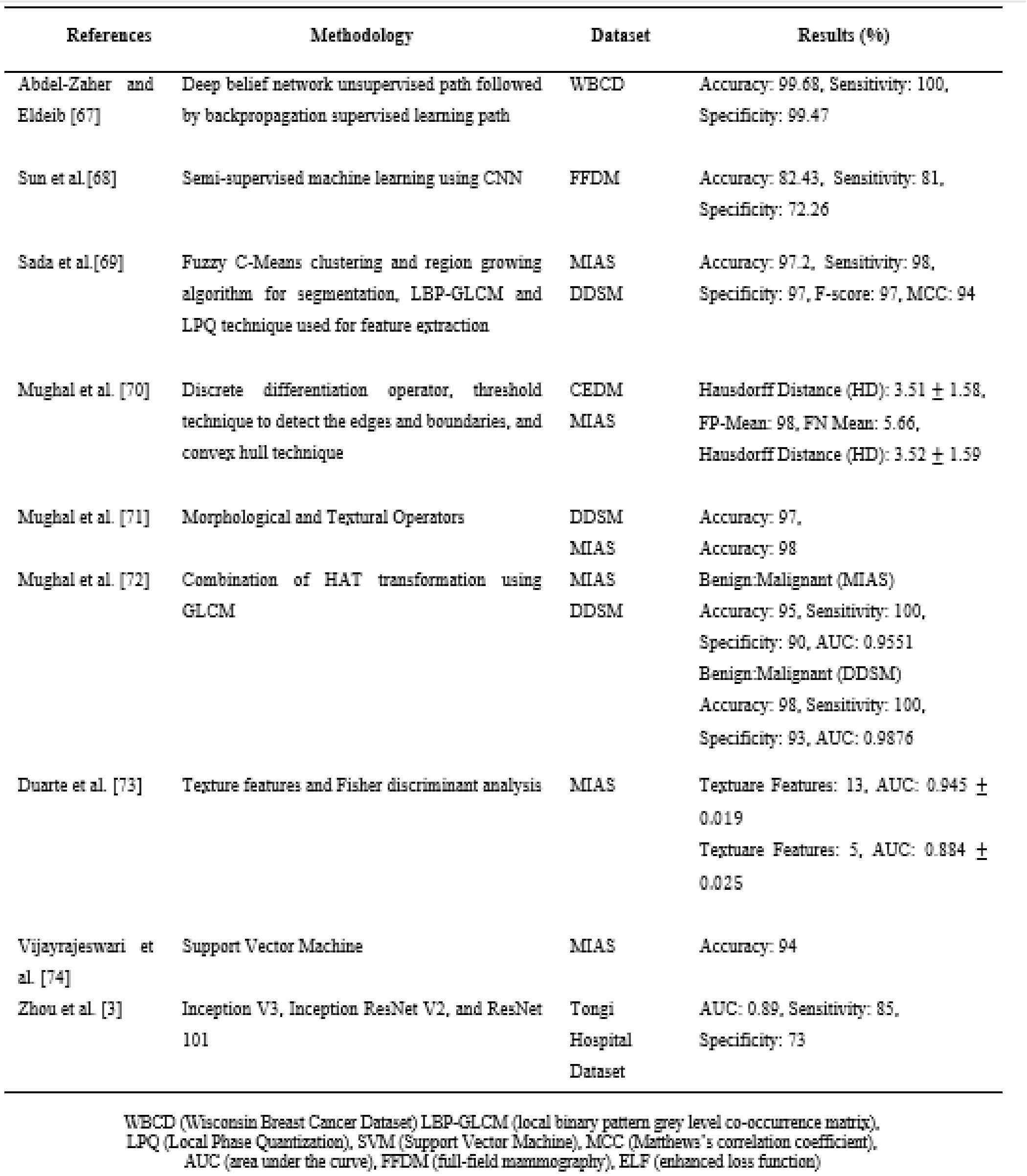


Figure 3.4: Work flow of a prediction model for cancer detection

Table 3.1: Current breast cancer detection technologies, datasets, and outcomes



References	Methodology	Dataset	Results (%)
Abdel-Zaher and Eldeib [67]	Deep belief network unsupervised path followed by backpropagation supervised learning path	WBCD	Accuracy: 99.68, Sensitivity: 100, Specificity: 99.47
Sun et al.[68]	Semi-supervised machine learning using CNN	FFDM	Accuracy: 82.43, Sensitivity: 81, Specificity: 72.26
Sada et al.[69]	Fuzzy C-Means clustering and region growing algorithm for segmentation, LBP-GLCM and LPQ technique used for feature extraction	MIAS DDSM	Accuracy: 97.2, Sensitivity: 98, Specificity: 97, F-score: 97, MCC: 94
Mughal et al. [70]	Discrete differentiation operator, threshold technique to detect the edges and boundaries, and convex hull technique	CEDM MIAS	Hausdorff Distance (HD): 3.51 ± 1.58, FP-Mean: 98, FN Mean: 5.66, Hausdorff Distance (HD): 3.52 ± 1.59
Mughal et al. [71]	Morphological and Textural Operators	DDSM MIAS	Accuracy: 97, Accuracy: 98
Mughal et al. [72]	Combination of HAT transformation using GLCM	MIAS DDSM	Benign:Malignant (MIAS) Accuracy: 95, Sensitivity: 100, Specificity: 90, AUC: 0.9551 Benign:Malignant (DDSM) Accuracy: 98, Sensitivity: 100, Specificity: 93, AUC: 0.9876
Duarte et al. [73]	Texture features and Fisher discriminant analysis	MIAS	Texuare Features: 13, AUC: 0.945 ± 0.019 Texuare Features: 5, AUC: 0.884 ± 0.025
Vijayrajeswari et al. [74]	Support Vector Machine	MIAS	Accuracy: 94
Zhou et al. [3]	Inception V3, Inception ResNet V2, and ResNet 101	Tongi Hospital Dataset	AUC: 0.89, Sensitivity: 85, Specificity: 73

WBCD (Wisconsin Breast Cancer Dataset) LBP-GLCM (local binary pattern grey level co-occurrence matrix), LPQ (Local Phase Quantization), SVM (Support Vector Machine), MCC (Matthews's correlation coefficient), AUC (area under the curve), FFDM (full-field mammography), ELF (enhanced loss function)

It is demonstrated the workflow of an existing prediction model for cancer detection. The major components are listed there. In between training model can varies preceded by the segmentation process but depends on the behavior of the data and user requirements. Finally, classification is done based on the input model as shown in Figure 3.5. Segmentation is the major process that helps to target the ROIs in the WSI samples. Many researchers developed supervised models to identify the ROI with the help of pre annotated data for training purposes [75][76][77]. Now, unsupervised models are also using to target the ROIs. Due to the limitation of computation capacity and memory space, the WSI cropped into smaller tiles then easy to do preprocessing and feature extraction. After predicting the label of each tile, finally merging will produce the overall segmentation and final prediction result of the WSI [44]. The list of methods, dataset, and results are demonstrated in Table 3.1.

3.5 Exploratory Analysis and Visualization

Pathology Predictive modelling has typically been the emphasis of imaging informatics. However, for two reasons, the study focus has shifted to a amalgamation of exploratory analysis and predictive modelling. For starters, large-scale studies like the TCGA aim to uncover new information regarding antagonistic cancer endpoints and identify new prognostic subtypes. Second, predictive modelling with high-dimensional data is extremely complex and necessitates the use of tools for assessing the biological applicability of characteristics as well as quantitative models. In addition, software applications known as "virtual microscopes" have arisen that allow for the spatial study of high-resolution digital WSI images.

3.6 Chapter Summary

This chapter, it is explained the model for a clinical decision support system that is an integrated part of Computer Aided Diagnosis. It is discussed the different biomarkers helpful for different cancer, quality control over digital slides, and how to handle the artifacts and other issues. It also included the quantitative description of an image at a pixel, object, and sematic level. In this chapter, the different predictive modeling and its visualizations aspects are discussed.

Chapter 4

A Novel Approach to Localize the ROI in Whole Slide Images

4.1 Introduction

Recent technological improvements in digital pathology and microscopy have been accomplished to support pathologists and clinicians in clinical diagnosis of cancer disease, reducing costs and increasing efficiency. The FDA of the US has released standards for the method, quality, and development of DICOM-compliant digital whole slide image scanning systems. The standard digital slides helping to the experts via CAD to diagnose and prognosis of different cancer. The whole slide images are supported by IHC, multispectral imaging concepts, and fluorescence. Identifying the ROI in an input sample image is usually a difficult operation due to the high resolution of WSI images, which necessitates good computation. The project aims to determine the optimal ROI using unsupervised machine learning and computational analysis of cancerous WSI sample pictures backed by shape formulas and morphological features. Because of the computational challenges, it's best to start with small patches, consolidate the data, and automatically localize the ROI. The work is when contrasted to automated ROI to the handcrafted ROI provided in the ICIAR 2018 dataset.

4.2 Literature Survey

4.2.1 Global Scenario of Breast Cancer

Due to different reasons, the number of cancer patients is increasing exponentially worldwide. World Cancer Report says in the year 2020, about 15 million additional cases are projected. Cancer appears to be tightening its grip over the world, particularly in developing and underdeveloped countries in

Asia, Africa, and South and Central America, as millions of new cases are reported each year. Breast cancer is one of the major cause of casualty among the women of all ages around the world [78]. Early identification and diagnosis of breast cancer, combined with therapy, significantly slows the disease's progression and lowers its morbidity rate [79].

4.2.2 Existing Techniques to Identify ROI using WSI

Whole Slide Imaging is a digitized microscopic image and one of the progressive fields of digital pathology. It inspects the different techniques and applied to upgrade cancer diagnosis and clinical care. Gilbertson and Wetzel formulated the first automated high-resolution WSI method in the history of digital pathology in 1999. Following numerous developments and adjustments in digital imaging hardware and methodologies, digital pathology practitioners have now incorporated these technical advancements and are steadily developing [60]. WSI provides virtual slides with a high resolution of less than 0.5m/pixels, which may be examined and inspected using interactive software on a good computer screen, in addition to the traditional pathology approach [80].

The WSI system makes great promises in the field of digital pathology. Similarly, various difficulties limit this concept, including inability to view the complete slide in high resolution, image quality, navigation control, excessive time to accurately study the slides, and adaptability to the system. Pathologists and clinicians are under pressure to improve patient safety, quality, and diagnosis accuracy with high clarity and precision. These factors prompted the manufacturer to create a system that would do better approach to multiple expert opinions and highly specialized pathological assistance while still providing a user-friendly interface. Digital pathology networks with virtual microscopes are a viable answer to all of these issues, and they will likely to play a key role in the future. Because it focuses primarily on the adaption of modern technologies,

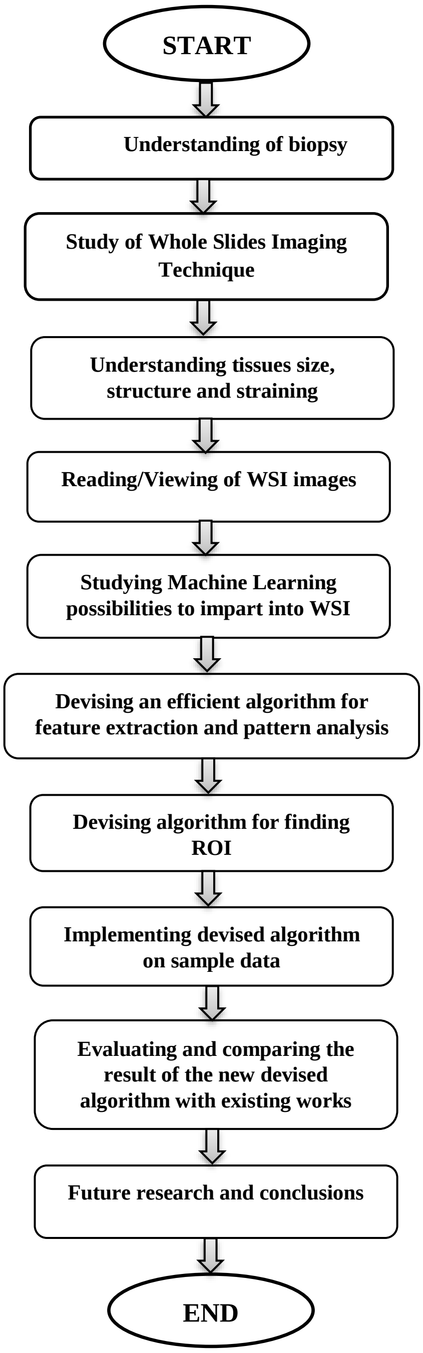

START
Understanding of biopsy
Study of Whole Slides Imaging Technique
Understanding tissues size, structure and straining
Reading/Viewing of WSI images
Studying Machine Learning possibilities to impart into WSI
Devising an efficient algorithm for feature extraction and pattern analysis
Devising algorithm for finding ROI
Implementing devised algorithm on sample data
Evaluating and comparing the result of the new devised algorithm with existing works
Future research and conclusions
END

Figure 4.1: Methodology of proposed work

many hospitals and healthcare organizations have accepted it. The authentication of the WSI system is described in the expanding research and literature review. However, finding the ROI is always a difficult operation due to gigapixel size and heavy computation. After training on picture patches rather than the entire image, a patch level classifier like CNN was constructed. Many similar efforts have been done in the last decade, with the majority of them relying on machine learning. It is recommended that the model be trained using CNN with picture patches of 500x500 pixels derived from big WSI datasets [44].

Following patch extraction, segmentation and likelihood in an Expectation-Maximization (EM) based approach are used to find eligible patches for training the CNN model. The WSI now allows pathologists and researchers to view digitized slides and obtain a better grasp of cancer diagnosis and decision-making processes. It was created a model that successfully detects the relevant ROIs with an accuracy of 74% using a a visual BoG and sliding window approach [75].

4.3 Materials and Methods

4.3.1 ICIAR 2018 WSI Data Samples

The ICIAR 2018 conference on BACH consists of histology microscope images stained with H&E samples from the entire slide [81]. There are 30 WSI accessible for training and 10 WSI available for model testing. All ten WSI malignant samples include pixel-wise annotated areas for the Benign, Invasive Carcinoma, and In Situ Carcinoma classifications, which have been labelled by pathologists and specialists. All of the slide photographs were captured using a Leica SCN400 in .svs format, in RGB color space, with a scale of 0.467m/pixels and an image size of (42113 x 62625) pixels. The python programming language (Python 3.6, 32-bit) is used to read the data and perform other tasks, and it is backed by open-source supporting library packages such as opencv, numpy, , imutils, sklearn, scikit-image, matplotlib, and openslide.

4.3.2 Proposed Methodology, Workflow, and Algorithm

The proposed methodology adopted a stepwise approach to proving the research task in Figure 4.1.

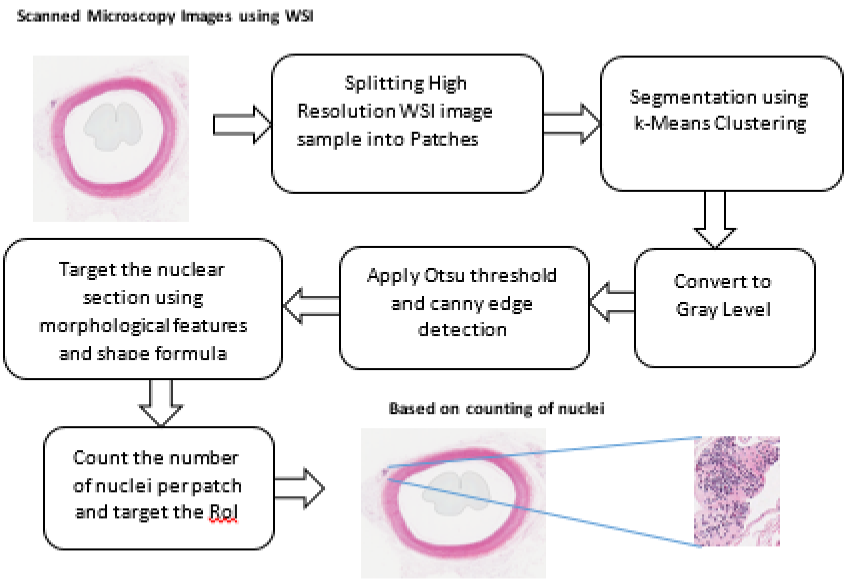

Scanned Microscopy Images using WSI
Splitting High Resolution WSI image sample into Patches
Segmentation using k-Means Clustering
Convert to Gray Level
Apply Otsu threshold and canny edge detection
Target the nuclear section using morphological features and shape formula
Count the number of nuclei per patch and target the RoI
Based on counting of nuclei

Figure 4.2: Workflow of the proposed methodology to localize ROI on WSI image

Algorithm 4.1: Targeting ROI Based on Localization and Counting of Nuclei

Input:

$\Re$: entire WSI image containing n objects,

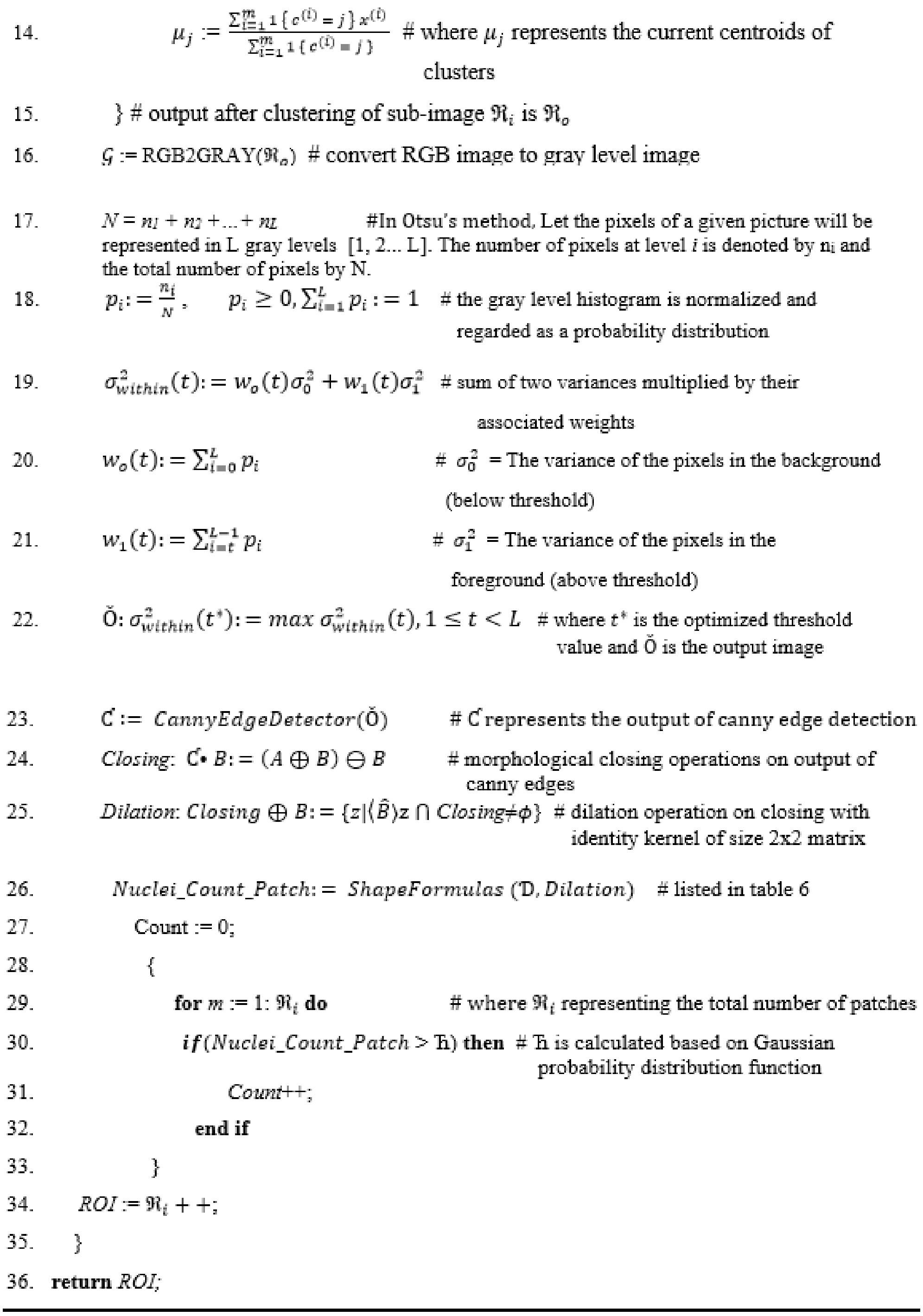
k: the number of clusters, $\mathcal{G}$: gray level image,

Ɖ: diameter of single nucleus given by pathologist (depends upon zoom level),

Ћ: Threshold nuclei count

Output: A set of the region of interest (ROI).

```
$\bigcup_{i=1}^{n} \Re_i := \Re$   # splitting the $\Re$ into sub-regions $\Re_1, \Re_2, \dots, \Re_n$
$\Re_i$ is a connected regions, $i := 1, 2, \dots, n$
$\Re_i \cap \Re_j := \emptyset$ for all i and j, i ≠ j
$P(\Re_i) := TRUE\ for\ i := 1, 2, \dots, n$   # where $P(\Re_i)$ is a logical predicate
$P(\Re_i \cup \Re_j) := FALSE\ for$ i ≠ j
Repeat for each sub-regions:
{
  arbitrarily choose the cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \Re_i$ as $k$ objects from $\Re$
    Repeat until convergence:
    {
      for every $i$, set
        $c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2$  # $||x^{(i)} - \mu_j||^2$ is the Euclidean distance
                                                   between $x^{(i)}$ and $\mu_j$
        for each $j$, set
```



```
                $\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$  # where $\mu_j$ represents the current centroids of clusters
      } # output after clustering of sub-image $\mathfrak{R}_i$ is $\mathfrak{R}_o$
    $\mathcal{G}$ := RGB2GRAY($\mathfrak{R}_o$)  # convert RGB image to gray level image
    $N = n_1 + n_2 + ... + n_L$    #In Otsu's method, Let the pixels of a given picture will be represented in L gray levels [1, 2... L]. The number of pixels at level $i$ is denoted by $n_i$ and the total number of pixels by N.
    $p_i := \frac{n_i}{N}$,   $p_i \geq 0, \sum_{i=1}^{L} p_i := 1$   # the gray level histogram is normalized and regarded as a probability distribution
    $\sigma^2_{within}(t) := w_o(t)\sigma_0^2 + w_1(t)\sigma_1^2$   # sum of two variances multiplied by their associated weights
    $w_o(t) := \sum_{i=0}^{L} p_i$    # $\sigma_0^2$ = The variance of the pixels in the background (below threshold)
    $w_1(t) := \sum_{i=t}^{L-1} p_i$    # $\sigma_1^2$ = The variance of the pixels in the foreground (above threshold)
    Ŏ: $\sigma^2_{within}(t^*) := max\ \sigma^2_{within}(t), 1 \leq t < L$   # where $t^*$ is the optimized threshold value and Ŏ is the output image
    Ć := $CannyEdgeDetector$(Ŏ)    # Ć represents the output of canny edge detection
    $Closing$: Ć• $B := (A \oplus B) \ominus B$    # morphological closing operations on output of canny edges
    $Dilation$: $Closing \oplus B := \{z | \langle \hat{B} \rangle z \cap Closing \neq \phi\}$  # dilation operation on closing with identity kernel of size 2x2 matrix
      $Nuclei\_Count\_Patch := ShapeFormulas$ (Ɗ, $Dilation$)   # listed in table 6
        Count := 0;
         {
           **for** $m$ := 1: $\mathfrak{R}_i$ **do**     # where $\mathfrak{R}_i$ representing the total number of patches
             $if$($Nuclei\_Count\_Patch$ > $\hbar$) **then**  # $\hbar$ is calculated based on Gaussian probability distribution function
                 $Count$++;
             **end if**
         }
    $ROI := \mathfrak{R}_i + +$;
  }
**return** $ROI$;
```

With the help of image segmentation and nucleus counting, feature extraction, and pattern analysis, all of the essential procedures are involved in detecting accurate ROI on WSI images as shown in Figure 4.2. The scanned WSI samples are given as input to the model and the final output is the targeted ROIs. All the steps are mathematically discussed in Algorithm 4.1.

4.3.3 Split WSI into Patches and Segmentation using k-Means Clustering

As an input, scanned microscope WSI pictures are used. Because of the high resolution of the WSI sample images, computing the entire image at once is quite complex. The sample image has a resolution of 15368 by 17496 pixels and three color channels. It is preferable to divide it into several patches. As a result, the sample image has been divided into 64 separate patches, each with a dimension of 1921 x 2187 as shown in Figure 4.3.

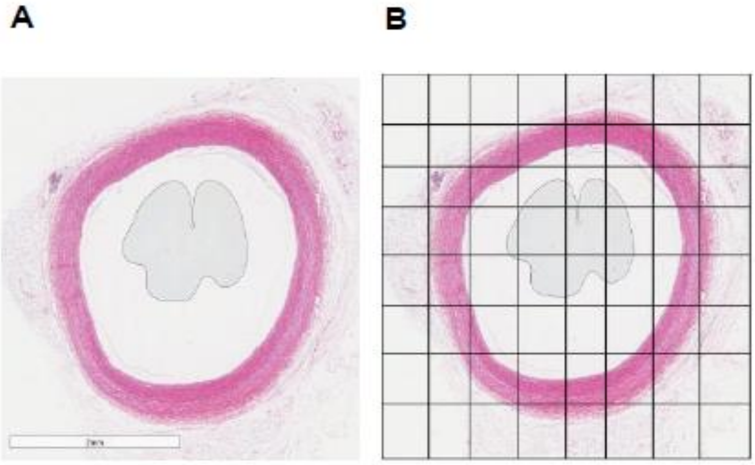
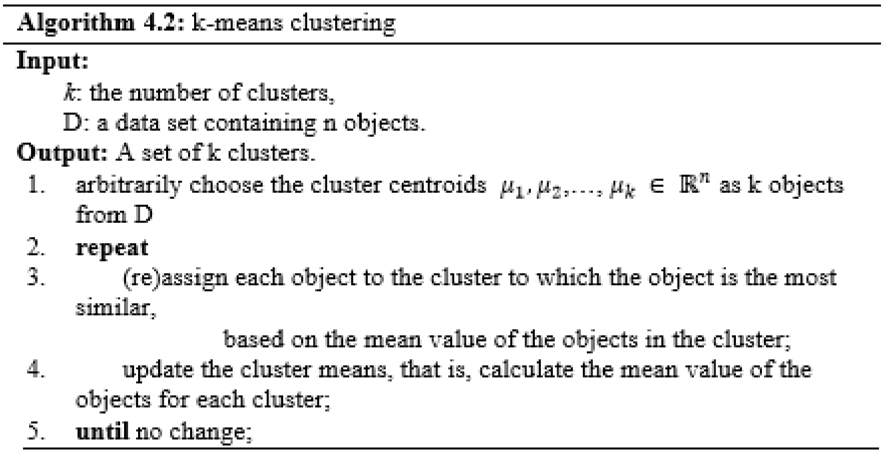


Figure 4.3: Split high-resolution WSI sample image (A) into low dimensional patches in (B)

Even yet, if the dimension is larger, it can be divided further until the scanner's true range of operation is reached. Every patch is now segmented using one of the most efficient unsupervised learning methods, k-means clustering as discussed in Algorithm 4.2 [40]. It smoothed the sample image and suppressed the abnormalities as sown in Figure 4.4.



```
Algorithm 4.2: k-means clustering
Input:
    k: the number of clusters,
    D: a data set containing n objects.
Output: A set of k clusters.
    arbitrarily choose the cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ as k objects
    from D
    repeat
        (re)assign each object to the cluster to which the object is the most
    similar,
                based on the mean value of the objects in the cluster;
        update the cluster means, that is, calculate the mean value of the
    objects for each cluster;
    until no change;
```

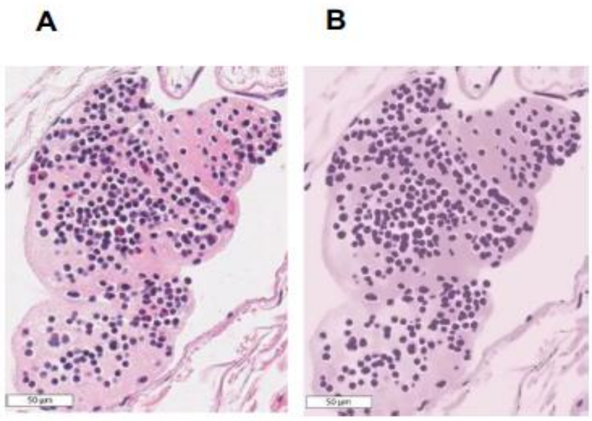


Figure 4.4: A) Implementation of k-means clustering algorithm on an image patch at a magnification level of x40 and B) results after clustering at the same magnification level

4.3.4 Usage of Otsu's Threshold and Canny Edge Detection Algorithm

After eliminating anomalies and smoothened the sample image, convert the image patches into an 8-bit gray-level image. Apply Otsu's thresholding method in the next stage [82][83]. It is always applied to the gray-level histogram and is one of the most effective approaches for obtaining threshold values. Let's say a given image pixels are depicted by L grey levels [1, 2... L]. Total number of pixels is calculated as N = $n_1 + n_2 + \ldots + n_L$, while the number of pixels at level *i* is denoted by n_i. The grey-level histogram is normalized and viewed as a probability distribution to simplify the discussion and written in Equation 4.1:

$$p_i = \frac{n_i}{N}, p_i \geq 0, \sum_{i=1}^{L} p_i = 1 \tag{4.1}$$

It depicts Otsu's method, which is within-class dissimilarity, represented as the sum of the two differences multiplied by their related weights, after applying the iterative methodology written in Equation 4.2.

$$\sigma_{within}^2(t) = w_o(t)\sigma_0^2 + w_1(t)\sigma_1^2 \tag{4.2}$$

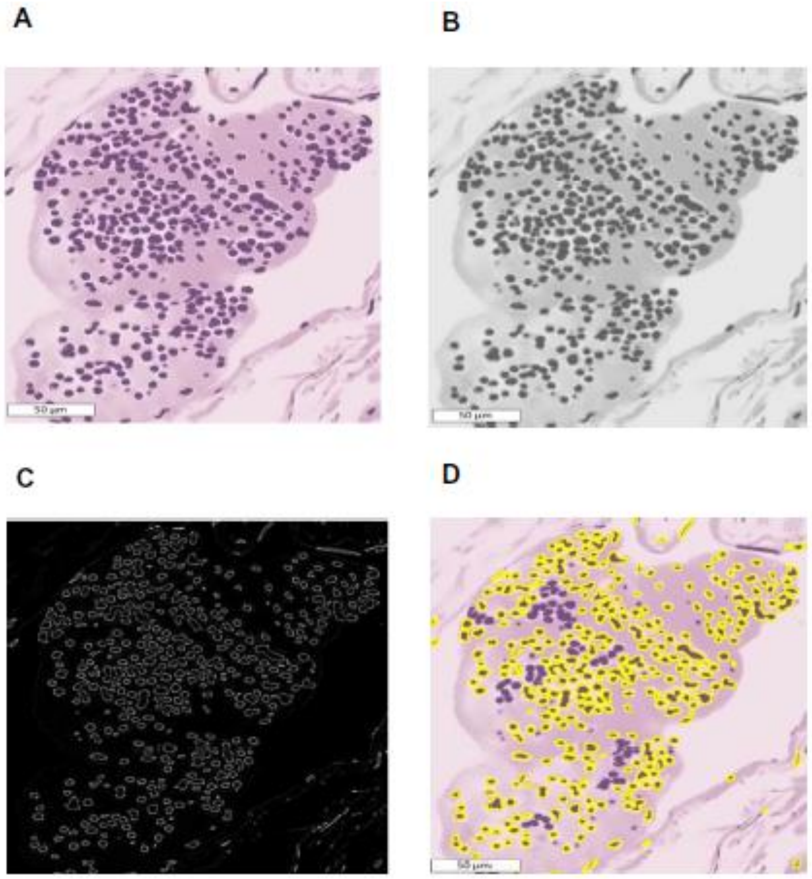


Figure 4.5: Implementation steps and the results at the magnification of x40 (A) represents the output of k-means clustering algorithm, (B) represents the 8-bit gray-level image, (C) represents the output of Otsu's method and canny edge detector, and (D) represents the localization of nuclei

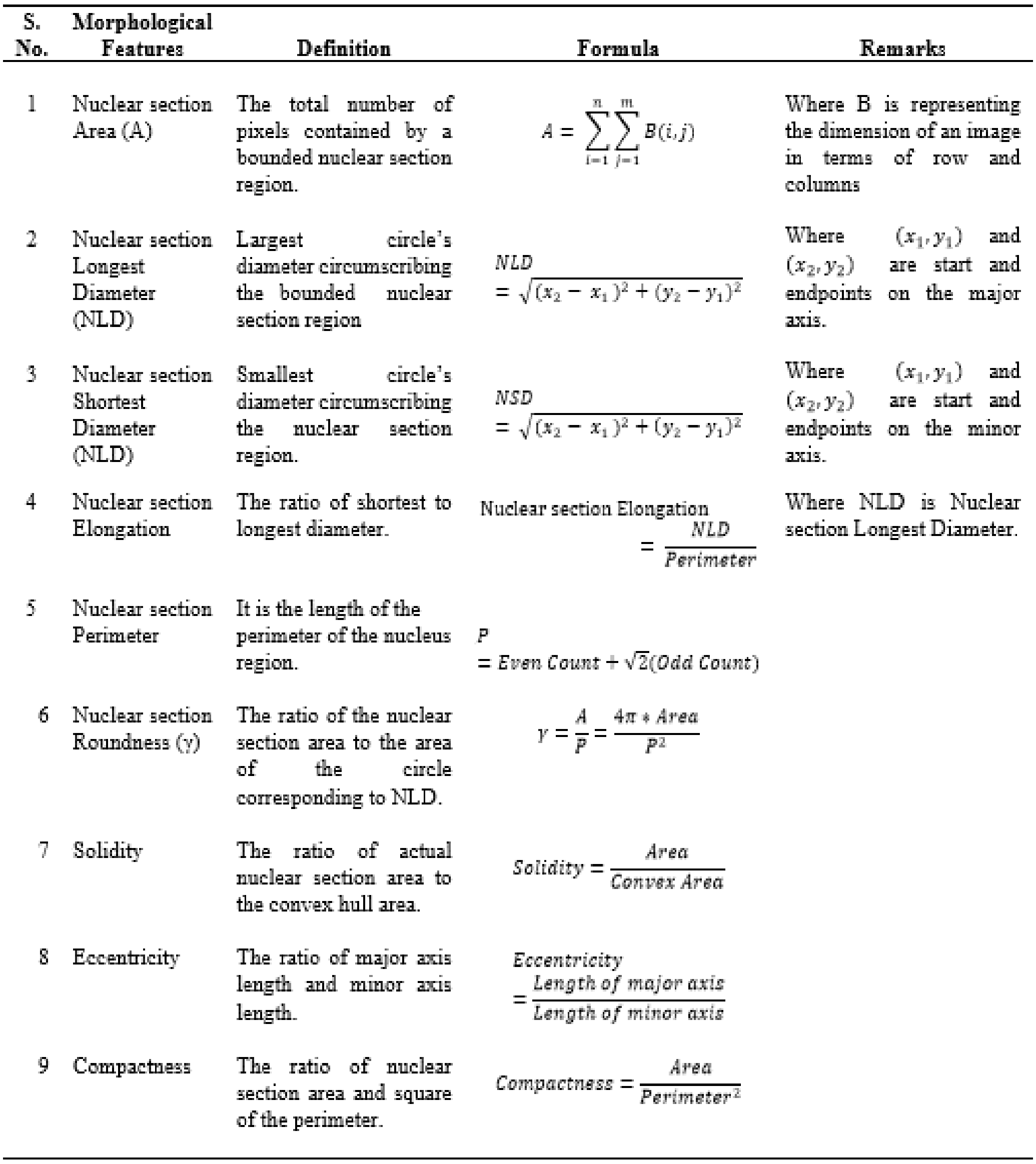
Where,

$$w_o(t) = \sum_{i=0}^{L} p_i, \qquad w_1(t) = \sum_{i=t}^{L-1} p_i$$

σ_0^2 = The below threshold value represents the abrupt changes of the pixels values to the background

σ_1^2 = The above threshold value represents the abrupt changes of the pixels values to the foreground

Table 4.1: Illustrating morphological shape and feature formulas



S. No.	Morphological Features	Definition	Formula	Remarks
1	Nuclear section Area (A)	The total number of pixels contained by a bounded nuclear section region.	$A = \sum_{i=1}^{n}\sum_{j=1}^{m} B(i,j)$	Where B is representing the dimension of an image in terms of row and columns
2	Nuclear section Longest Diameter (NLD)	Largest circle's diameter circumscribing the bounded nuclear section region	$NLD = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	Where (x_1, y_1) and (x_2, y_2) are start and endpoints on the major axis.
3	Nuclear section Shortest Diameter (NLD)	Smallest circle's diameter circumscribing the nuclear section region.	$NSD = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	Where (x_1, y_1) and (x_2, y_2) are start and endpoints on the minor axis.
4	Nuclear section Elongation	The ratio of shortest to longest diameter.	$\text{Nuclear section Elongation} = \frac{NLD}{Perimeter}$	Where NLD is Nuclear section Longest Diameter.
5	Nuclear section Perimeter	It is the length of the perimeter of the nucleus region.	$P = Even\ Count + \sqrt{2}(Odd\ Count)$	
6	Nuclear section Roundness (γ)	The ratio of the nuclear section area to the area of the circle corresponding to NLD.	$\gamma = \frac{A}{P} = \frac{4\pi * Area}{P^2}$	
7	Solidity	The ratio of actual nuclear section area to the convex hull area.	$Solidity = \frac{Area}{Convex\ Area}$	
8	Eccentricity	The ratio of major axis length and minor axis length.	$Eccentricity = \frac{Length\ of\ major\ axis}{Length\ of\ minor\ axis}$	
9	Compactness	The ratio of nuclear section area and square of the perimeter.	$Compactness = \frac{Area}{Perimeter^2}$	

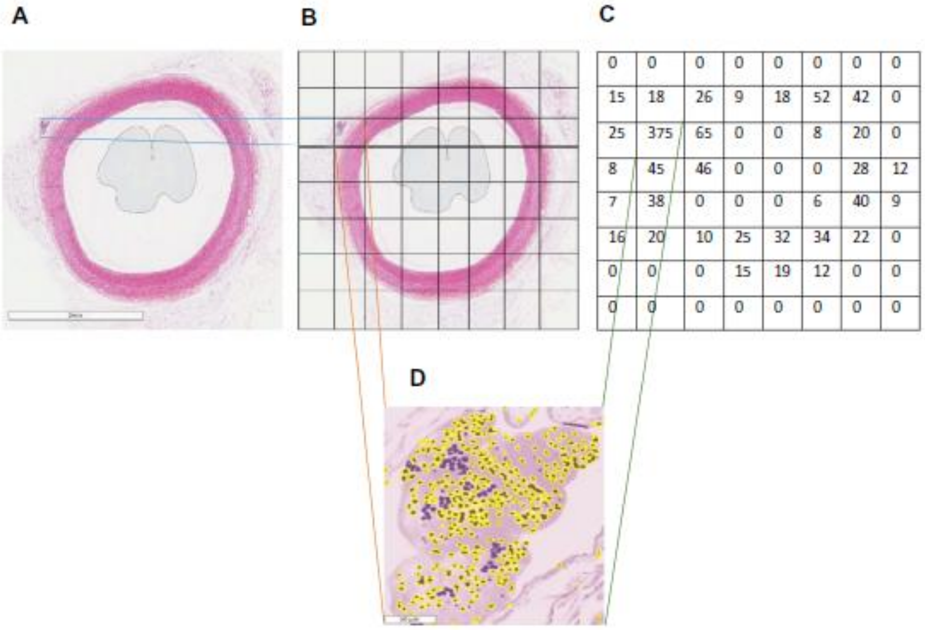
The global threshold value and edges in patches were found using Otsu's technique and the Canny edge detector, respectively as shown in Figure 4.5C. The discriminative ROI on sample data can be found using morphological features and different shape algorithms. Various textural and morphological properties such as nuclear section area, solidity, roundness, compactness, and so on play a critical role in identifying the nuclear section of sample photos. Experts (pathologists) start the process of determining the approximate size of the nuclear section; after that, the system takes over and detects the nuclei automatically. Finally, it will calculate the effective and approximate nuclei counts for each patch. It is simple to define the proper ROI on the whole WSI sample using nuclei counting.

In order to continue with the final result, the output of k-means clustering is turned into an 8-bit gray-level image as shown in Figure 4.5B. The Canny edge detection technique has been built after utilizing Otsu's method for detecting acceptable edges as shown in Figure 4.5C. The nuclei in the patch must be targeted and accurately counted as the next most essential step. Nucleus segmentation and localization are demonstrated in Figure 4.5D.

The morphological shapes and their formulas are listed in Table 4.1 that will help to calculate the parameters like nuclei area, perimeter, solidity, convexity, elongation, roundness, and compactness.

4.3.5 Target the ROI Based on Counting of Nuclei / Patch

The steps to target and localize the ROI are represented, as well as the results are shown in Figure 4.6. It is required the WSI sample images as an input. .



0	0	0	0	0	0	0	0
15	18	26	9	18	52	42	0
25	375	65	0	0	8	20	0
8	45	46	0	0	0	28	12
7	38	0	0	0	6	40	9
16	20	10	25	32	34	22	0
0	0	0	15	19	12	0	0
0	0	0	0	0	0	0	0

Figure 4.6: Steps and results of targeting ROI (A) high-resolution H&E stained sample image, B) split it into equal low dimension patches, (C) counting of nuclei present in each patch, and (D) represents pixel-wise points of interest (nuclei) at a magnification level of x40

Figure 4.6A shows the input sample image, which is divided into 64 equal patches as shown in Figure 4.6B. Figure 4.6C shows the quantitative representation of each low-resolution patch from Figure 4.6B. The nuclei counting of a targeted patch is observed to be at a maximum of 375 of the patch location (1, 2), as illustrated in Figure 4.6D. The number of nuclei in the other regions is extremely low. The threshold frequency is determined using statistical calculations such as mean, standard deviation, and variance. As a result, the patch located at (1, 2) will be most suited to become the ROI.

4.4 Experimental Process and Results

This section has described the novel approach to localize the ROI. To demonstrate the work, there are important steps that to follow. It started with reading high-dimensional WSI images, image viewing with the different zoom levels, panning, and creating digital slide repositories. In the second step, devising the efficient algorithm for feature extraction and its results for pattern analysis. In the third step, after working on different patches and the pattern analysis of each patch, it will be a localized region of interest (ROI) supported by shape and morphological operators on high dimensional images with achieved accuracy of 85.5% [84]. After localizing of ROI, the performance analysis will be done based on different quantitative accuracy measures.

Experimental setup

It is used HP workstation with i3 processors and 4GB of RAM. Based on experience, it is recommended to our research community not to use less than 4GB of RAM. It is listed the system configuration, programming language, and supporting python library packages in Table 4.2. It is targeting this research work for breast cancer. This effort will assist in locating the desired ROI. The snapshots of the developed tools and the results are given below to display and process the high-resolution images. It is used the pyramid concept for developing this tool. This tool is compatible with extensions of ndpi, svs, tiff, jpeg, etc.

Table 4.2: System configuration and Open Source tools used

Requirement	Specifications
Operating System	Windows 10, 64-bit
Processor	Core i3, 2.00GHz
Installed memory (RAM)	4 GB
Hard Disk Memory	100GB
Programming Language	Python, Java, JavaScript
Library Packages	OpenSlide, OpenLayer, Numpy, Sklearn, Matplotlib, OpenCV etc.

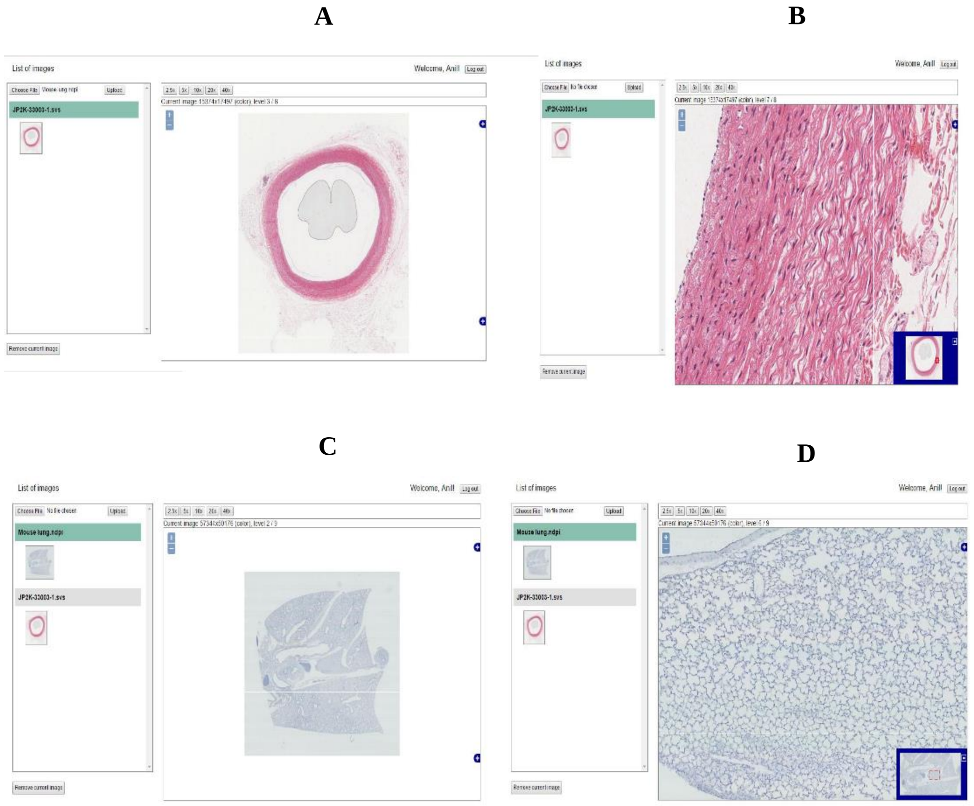
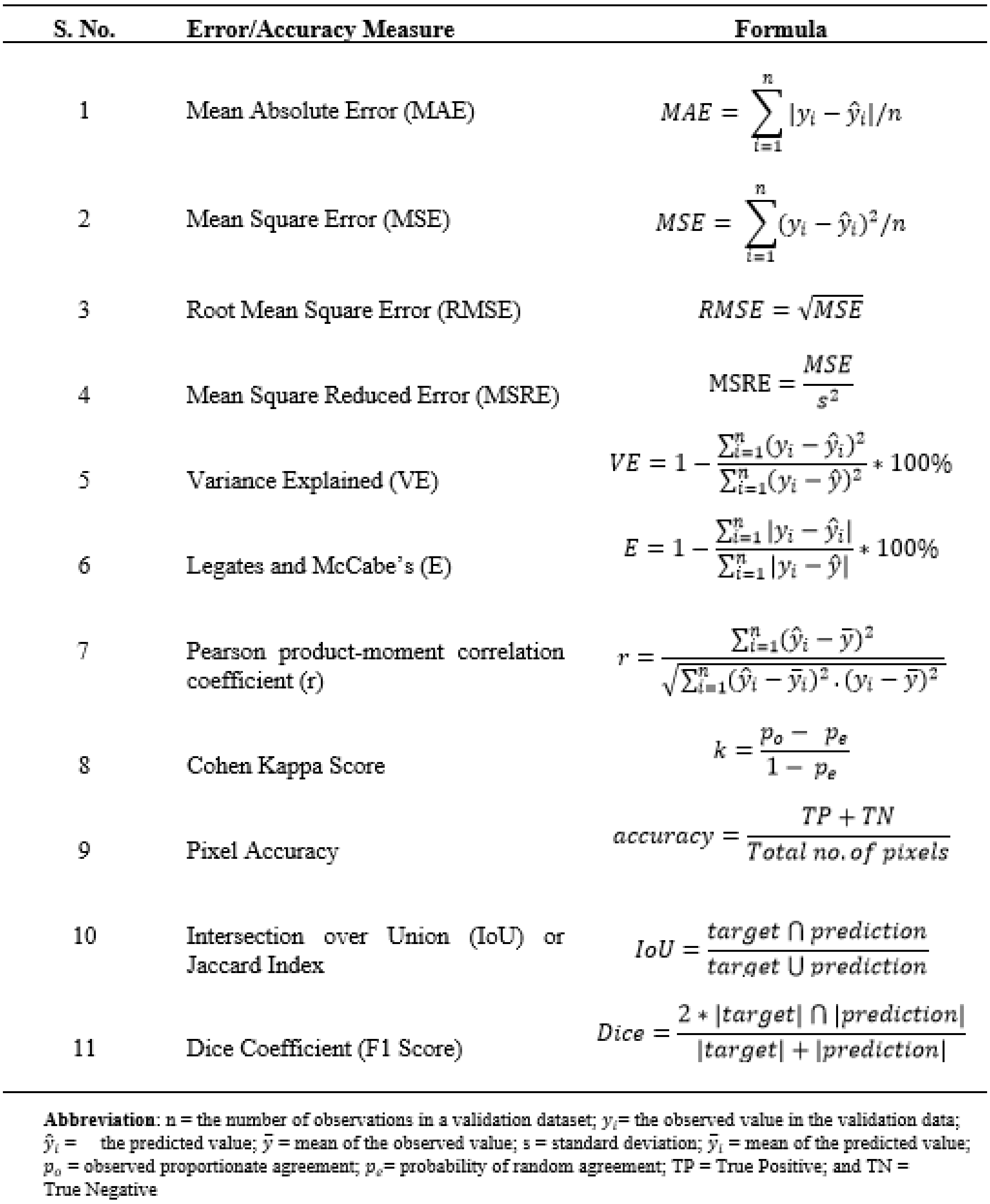


Figure 4.7: All the above snaps (A), (B), (C), and (D) are the samples repository and respective zoomed images. It will help pathologist to diagnose on high resolution screen

It is also used open layers packages. To create a directory, it is used JSON and javascript are embedded in HTML for web pages. This is very helpful to display the whole slide images. After analysis of such images on a high definition display monitor, the pathologist able to diagnosis different types of cancers. This tool is capable of zoom an image up to x40 and more as per the requirement and quality of the input image. It can add multiple images from various formats in the list that is showing on the left side of the tool shown in Figure 4.7.

Table 4.3: Relevant measures' mathematical definition/formula

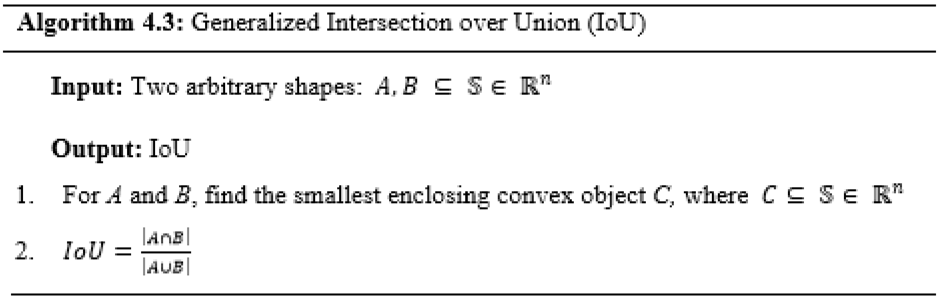


S. No.	Error/Accuracy Measure	Formula
1	Mean Absolute Error (MAE)	$MAE = \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert / n$
2	Mean Square Error (MSE)	$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n$
3	Root Mean Square Error (RMSE)	$RMSE = \sqrt{MSE}$
4	Mean Square Reduced Error (MSRE)	$\mathrm{MSRE} = \frac{MSE}{s^2}$
5	Variance Explained (VE)	$VE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} * 100\%$
6	Legates and McCabe's (E)	$E = 1 - \frac{\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert}{\sum_{i=1}^{n} \lvert y_i - \hat{y} \rvert} * 100\%$
7	Pearson product-moment correlation coefficient (r)	$r = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 \cdot (y_i - \bar{y})^2}}$
8	Cohen Kappa Score	$k = \frac{p_o - \ p_e}{1 - \ p_e}$
9	Pixel Accuracy	$accuracy = \frac{TP + TN}{Total\ no.of\ pixels}$
10	Intersection over Union (IoU) or Jaccard Index	$IoU = \frac{target \cap prediction}{target \cup prediction}$
11	Dice Coefficient (F1 Score)	$Dice = \frac{2 * \lvert target \rvert \cap \lvert prediction \rvert}{\lvert target \rvert + \lvert prediction \rvert}$

Abbreviation: n = the number of observations in a validation dataset; y_i= the observed value in the validation data; $\hat{y}_i$ = the predicted value; $\bar{y}$ = mean of the observed value; s = standard deviation; $\bar{y}_i$ = mean of the predicted value; p_o = observed proportionate agreement; p_e= probability of random agreement; TP = True Positive; and TN = True Negative

The BACH dataset, a collection of breast cancer histology image samples are used to validate the findings. WSI samples of H&E stained breast histology microscope images make up the BACH dataset. Only ten pixel-wise labelled WSI cancerous samples are being studied in this study. Many pathologists and medical specialists are the members of the ICIAR 2018 organizing team and helped to label and annotate all of the available 10 samples. Each image has three classifications: benign, in-situ cancer, and invasive carcinoma. The image's remaining unnamed area will be regarded normal. Before beginning the work, all three kinds of benign, in-situ carcinoma, and invasive carcinoma are grouped into one category termed affected area or targeted ROIs, which is shown by red, while the rest (background) is represented by black.

The similarity index of the actual and projected results is used to calculate the work's accuracy measurements. The relevance of intersection over union is highlighted (IoU). It is one of the most often used object detection benchmarks. The overlapping notion is used to compare the similarity of two arbitrary forms using the IoU metric that derived in Algorithm 4.3 [85].



Algorithm 4.3: Generalized Intersection over Union (IoU)

```
Input: Two arbitrary shapes: $A, B \subseteq \mathbb{S} \in \mathbb{R}^n$
Output: IoU
1. For $A$ and $B$, find the smallest enclosing convex object $C$, where $C \subseteq \mathbb{S} \in \mathbb{R}^n$
2. $IoU = \frac{|A \cap B|}{|A \cup B|}$
```

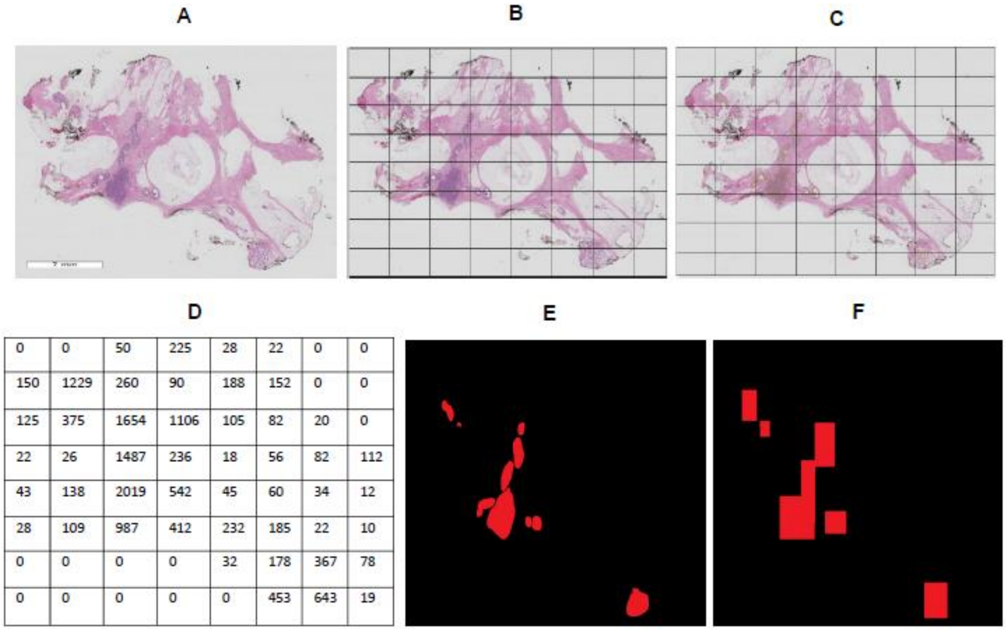
It is listed the mathematical formula for relevant measures in Table 4.3 that are finally calculated and listed in Table 4.4. Different accuracy measures are compared with IoU in Table 4.4. This research will aid pathologists in accurately diagnosing cancer patients with greater clarity and precision in less time. The majority of a pathologist's time is spent diagnosing sample tissues. It takes so much time and effort due to the complexity of visualization and various ROIs. To define these ROIs and train probabilistic classifiers that help predict similar ROI on WSI samples, the visual BoG model with texture and colour features was implemented. This study included 240 WSI breast biopsies from 5 various degrees of malignancy, ranging from benign to malignant. And achieving a 79.8% accuracy rate in locating the correct ROI [75].

Another study used a rapid segmentation method combined with an instinctive multiclass supervised classification to construct a map of a WSI and spotted biological ROIs. Expert information was conveyed as morphological annotations [86]. A deep learning algorithm is also used to discuss the entire slide cancer diagnosis. It presented a strategy for mastering the potential to automate subject expert like diagnostic ability and convert the gigapixels straight into a series of fine predictions, providing multiple opinions, and fostering clinical pathology consensus [87]. It is applied the v3 DCNN model and generating a FROC of 83.5 percent. It is employed in this research. For clinicians, the Camelyon16 dataset automatically generates a WSI heatmap and extracts polygons of lesion areas [88].

The ICIAR 2018 BACH WSI dataset is used in all of the studies below. An ensemble of CNN is used to offer an automated classification approach for recognizing the microstructures of tissues that has a 55.26% of accuracy[89]. In microscope and WSI annotated data set, a classification and localization strategy for clinically relevant histopathological classes is proposed. The presented technique was an upgraded version of state-of-the-art CNN model that achieved an average accuracy of 69% for automatically recognizing and classifying the ROI [50]. By using smart tactics like mirroring, rotating, and fine-tuning of pre-trained networks, it has been attempted to lower the cost of collecting medical data.

In order to continue this effort, a DCNN (ALEXNET) was fine-tuned and reached an average accuracy of 75.73% [90]. After fine-tuning the Inception-v3 the CNN algorithm is proposed. It extracts patches based on nuclear density and eliminates areas with a low number of nuclei. Every patch with a high nuclear density is accepted, and the nuclear classes are defined with an average accuracy of 79& based on majority voting [91]. For automatic classification of the WSI dataset, a patch-based classifier based on CNN is presented. The patch-based classifier predicts the class label of each patch using OPOD, and then uses majority voting procedures to make a final judgement on the WSI sample image's final class label. The algorithm's average patch-wise classification accuracy is 81.05% [92]. For the categorization of breast cancer histopathology samples, a novel hybrid convolutional and recurrent DNN is presented.

The approach is based on multilayer feature representation and incorporates advantages of CNN and RCNN. The spatial association between patches, both short-term and long-term, is preserved. For the typical class, it achieved an average accuracy of 82.1% [93].



0	0	50	225	28	22	0	0
150	1229	260	90	188	152	0	0
125	375	1654	1106	105	82	20	0
22	26	1487	236	18	56	82	112
43	138	2019	542	45	60	34	12
28	109	987	412	232	185	22	10
0	0	0	0	32	178	367	78
0	0	0	0	0	453	643	19

Figure 4.8: Steps implemented on BACH high resolution WSI sample images (A) H&E stained malignant sample image, (B) split it into low dimension, (C) target the patch based on highest counting of nuclei (D) using algorithm, count the number of nuclei for each patch, (E) manually annotated by experts, and (F) automated annotation of target patches.

It is demonstrated one sample WSI of the ICIAR 2018 BACH dataset in Figure 4.8. After executing all the steps of Figure 4.2, it found the automated annotation of the predicted patches. These predicted patches are based on the counting of nuclei for each segment. The predicted patches are finally validated by corresponding manually annotated patches that are done by experts.

This work is done on the ICIAR 2018 conference on breast cancer histology specimen to validate the results. The H&E stained breast histology microscopic slides are scanned and archived as a collection. Although there are several WSI data samples but this study focuses only on 10 pixel-wise labelled WSI.

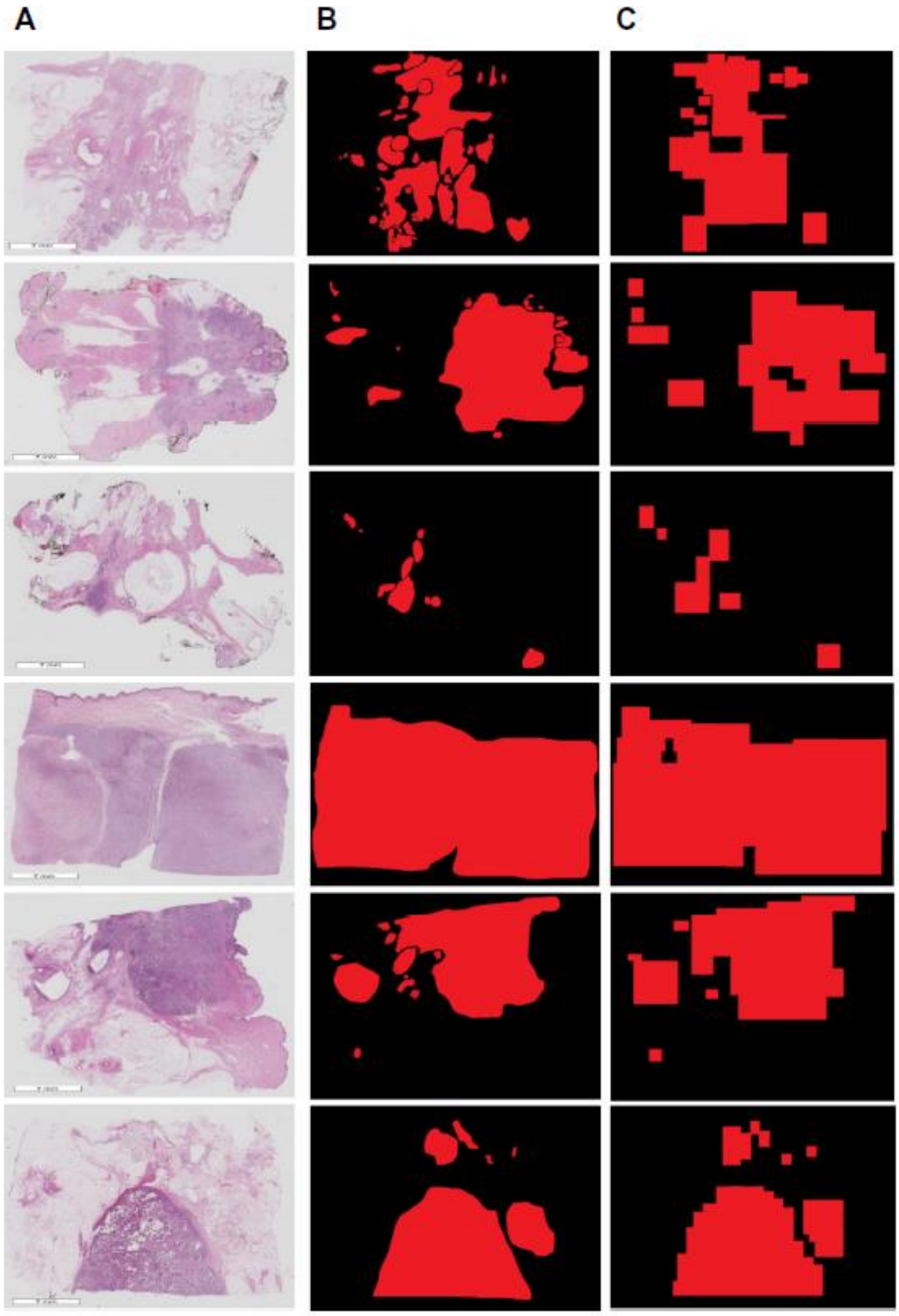
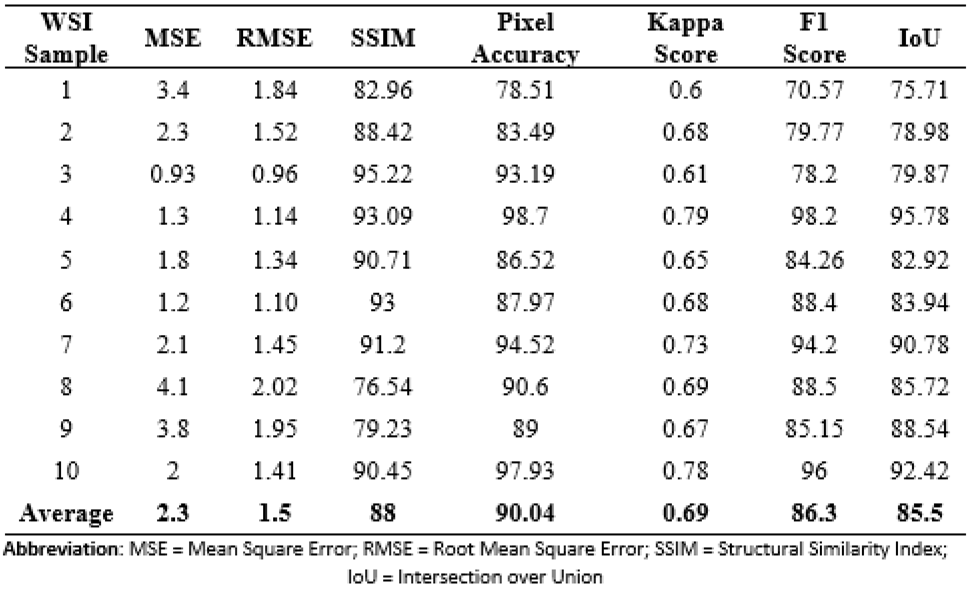


Figure 4.9: Targeted ROI on WSI image samples (A) represents WSI malignant samples of breast cancer in .svs format with a pixel scale of 0.467μm/pixels, (B) represents labeled and annotated by medical experts of the BACH challenge, and (C) represents automated localized ROI based on counting of nuclei

cancerous examples. Only six WSI image samples in Figure 4.9A are shown in the results section. Two medical specialists who are members of the organizing team who are responsible to label and annotate all the available 10 samples. Each image has three classifications: benign, in-situ, and invasive carcinoma. The image's remaining unnamed area will be regarded normal. Before beginning work on this article, all three types of benign, in-situ, and invasive carcinoma are grouped into one category termed affected area or targeted ROIs, which is represented by red, while the rest is represented by black. The expert's advice is followed for all of the goals shown in Figure 4.9B and finally, the ROI is predicted and shown in Figure 4.9C.

Table 4.4: Comparison of different accuracy measure implemented on proposed segmentation results



WSI Sample	MSE	RMSE	SSIM	Pixel Accuracy	Kappa Score	F1 Score	IoU
1	3.4	1.84	82.96	78.51	0.6	70.57	75.71
2	2.3	1.52	88.42	83.49	0.68	79.77	78.98
3	0.93	0.96	95.22	93.19	0.61	78.2	79.87
4	1.3	1.14	93.09	98.7	0.79	98.2	95.78
5	1.8	1.34	90.71	86.52	0.65	84.26	82.92
6	1.2	1.10	93	87.97	0.68	88.4	83.94
7	2.1	1.45	91.2	94.52	0.73	94.2	90.78
8	4.1	2.02	76.54	90.6	0.69	88.5	85.72
9	3.8	1.95	79.23	89	0.67	85.15	88.54
10	2	1.41	90.45	97.93	0.78	96	92.42
Average	**2.3**	**1.5**	**88**	**90.04**	**0.69**	**86.3**	**85.5**

Abbreviation: MSE = Mean Square Error; RMSE = Root Mean Square Error; SSIM = Structural Similarity Index; IoU = Intersection over Union

It is implemented the accuracy measures MSE, RMSE, SSIM, Pixel Accuracy, Kappa Score, F1 Score, and IoU are mentioned in Table 4.3. The corresponding average calculated values are 2.3%, 1.5%, 88%, 90.04%, 0.69, 86.3%, and 85.5% respectively for total of 10 WSI samples as shown in Table 4.4.

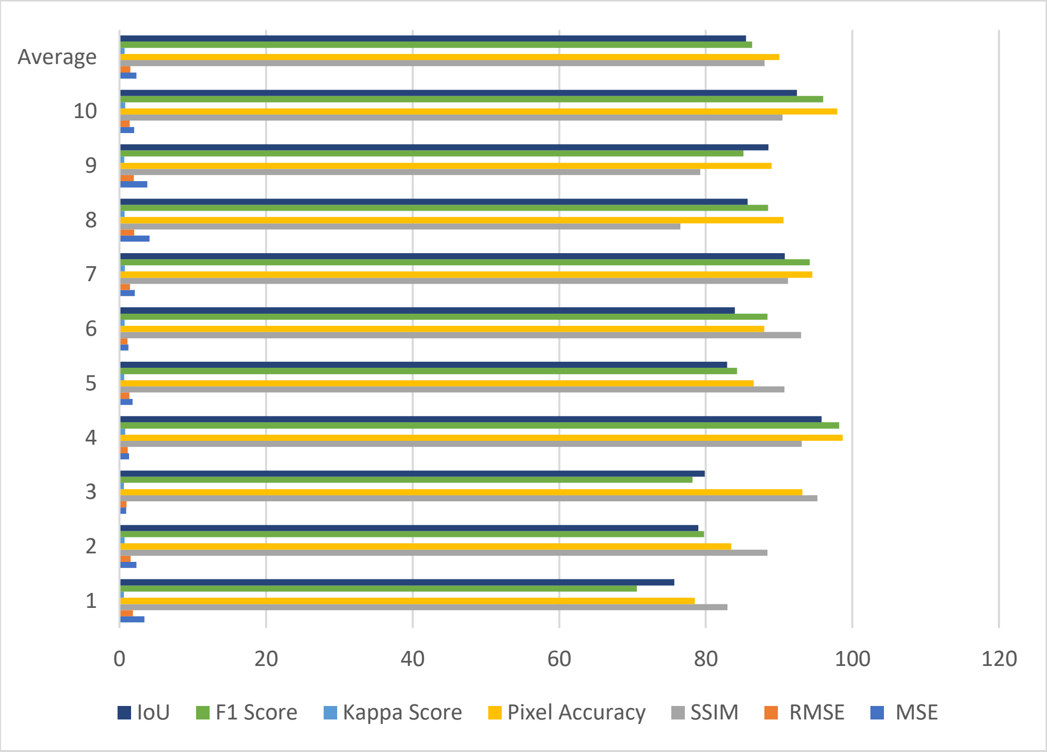
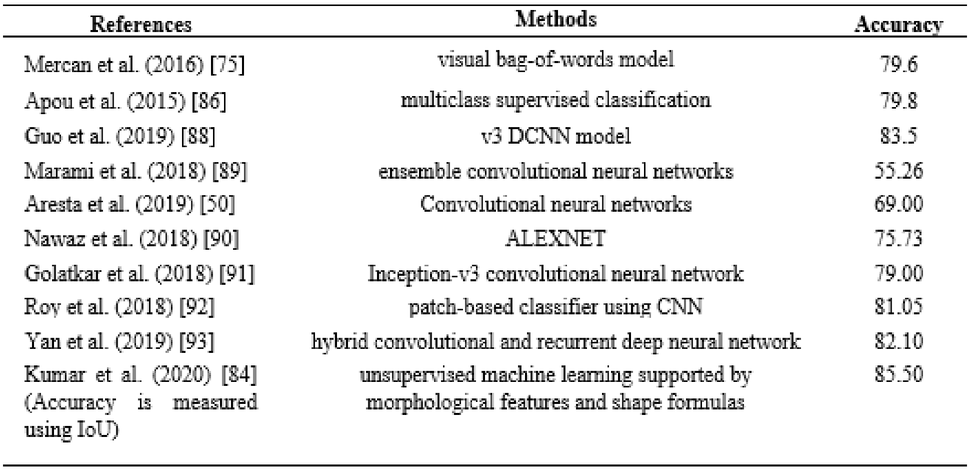


Figure 4.11: Graphical representation of comparison of different accuracy measure implemented on proposed segmentation results

Table 4.5: Quantitative accuracy comparison of different method vs this research work



References	Methods	Accuracy
Mercan et al. (2016) [75]	visual bag-of-words model	79.6
Apou et al. (2015) [86]	multiclass supervised classification	79.8
Guo et al. (2019) [88]	v3 DCNN model	83.5
Marami et al. (2018) [89]	ensemble convolutional neural networks	55.26
Aresta et al. (2019) [50]	Convolutional neural networks	69.00
Nawaz et al. (2018) [90]	ALEXNET	75.73
Golatkar et al. (2018) [91]	Inception-v3 convolutional neural network	79.00
Roy et al. (2018) [92]	patch-based classifier using CNN	81.05
Yan et al. (2019) [93]	hybrid convolutional and recurrent deep neural network	82.10
Kumar et al. (2020) [84] (Accuracy is measured using IoU)	unsupervised machine learning supported by morphological features and shape formulas	85.50

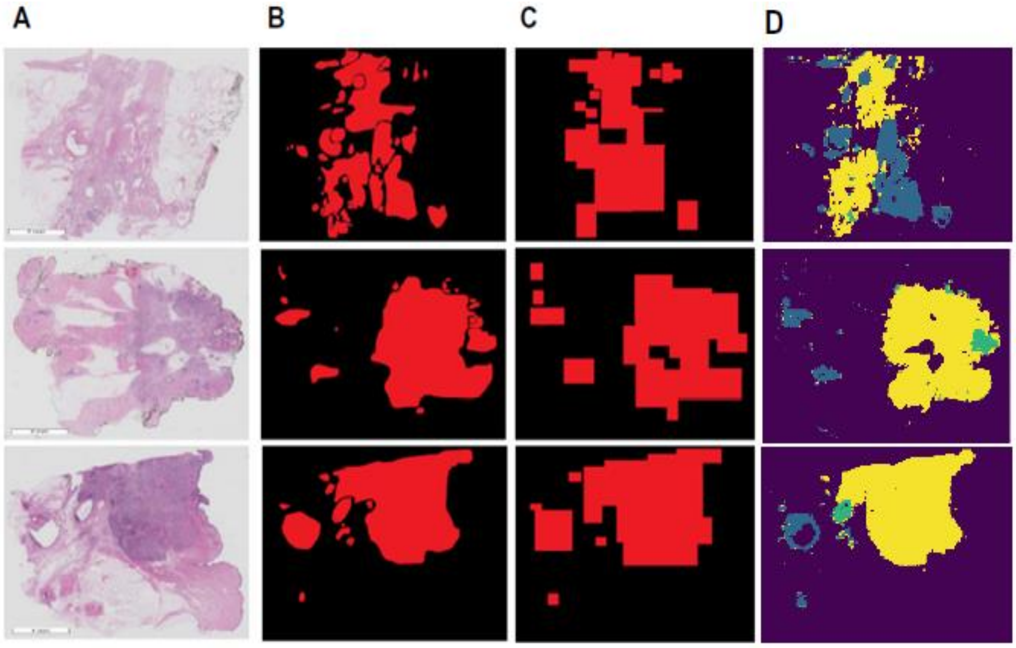


Figure 4.12: Comparison of the results of this work with one of the state-of-the-art work (A) represents WSI malignant samples of breast cancer in .svs format with a pixel scale of 0.467μm/pixels, (B) represents labeled and annotated by medical experts of the BACH challenge, (C) represents this work of automated localized ROI based on counting of nuclei, and (D) represents the results of Ensemble Network supported by ResNet classifier to region identification

It is applied Algorithm 4.1 to automatic identification of ROI on WSI samples. It is accepted the IoU based average accuracy of 85.5% which can be projected as a more optimized result in comparison to other methods listed in Table 4.5. The graphical representation of the segmentation accuracy of Table 4.4 is shown in Figure 4.10. To compare the visual difference of segmentation accuracy, it is illustrated the comparisons of the output of the proposed algorithm and ResNet classifier that is one of the state-of-the-art algorithms shown in Figure 4.11.

4.5 Chapter Summary

This chapter is describing a novel approach to localize the ROI in WSI, supported by shape formulas and morphological features on ICIAR 2018 BACH dataset. It is discussed the existing model to identify the ROI on WSI sample images. The proposed methodology, its framework and algorithm are explained supported by results and limitations.

Chapter 5

Prognostic Evaluation and Grading of Breast Cancer Using Ki-67 Antigen

This research work aimed significance of the Ki-67 antigen biomarker and computed the proliferation score based on the counting of immunopositive and immunonegative nuclear sections with the help of machine learning to predict the grading of breast cancer.

5.1 Literature Survey on Ki-67 Antigen and Biomarker

The Ki-67 antigen is a nuclear protein used as a cellular biomarker for breast cancer proliferation and is widely used in immunohistochemistry (IHC). It is observed on recent data, the grading of Ki-67 above the range 10%-14% mark out a high-risk category of cancer in terms of prognosis [94]. At the St Gallen Consensus in 2009, the expert panel suggested that the labeling index of Ki-67 is significant for selecting the treatment using radiotherapy and chemotherapy. The values of Ki-67 suggest grading the tumors as low, medium, and high proliferated based on the proliferation score of ⩽ 15%, 16%-30%, and > 30%, respectively [95]. Many studies have been conducted to assess the routine use and utility of Ki-67 as a prognostic grading index marker in breast cancer, with the goal of improving clinical care and avoiding needless chemotherapy. The "International Ki-67 in Breast Cancer Working Group" published their contribution based on current evidence in the areas of Ki-67 evaluation and quantifiable description, research to establish strong inter-laboratory chemistry, and scientific acceptance of the marker in clinical practice as one of the strong biomarkers studied and measured by immunohistochemistry (IHC) [5]. Because of insufficient quality commitments, the "American Society of Clinical Oncology (ASCO) Tumor Marker Guidelines Committee" did not recommend using Ki-67 for prognosis with newly diagnosed breast cancer in the beginning

[6]. The same is suggested in the original article “Ki-67 as a prognostic marker according to breast cancer molecular subtype” [9].

After working on a total of 107 selected cases of invasive breast cancer that the proliferation score of Ki-67 may be treated as a successful biomarker and it can be used for the treatment. The ki-67 grading index is a strong biomarker to contemplate neoadjuvant chemotherapy [96]. Patients with higher cell proliferation scores like Ki-67 > 25% may be treated by neoadjuvant chemotherapy. It is advised after studying breast cancer among 92 cases that the Ki-67 grading index manifested by immunohistochemical methods may be recognized as a potential biomarker and it can provide prognostic statistics like pathological tumor grading, size, and lymph node connection to decide benign or malignant tumor [97].

After analyzing all the above descriptions and different techniques, it is observed that segmentation is still one of the competitive tasks especially when there are high dimension images in digital pathology. Digital pathology involved the steps in which histology slides are digitized and generate high-dimensional images with the help of high configured whole slide digital scanners. The segmentation methods help to find the ROIs. The image analysis tasks can be performed after segmentation like capturing cell nuclear sections, tissue grading (classification), differentiating cancerous and noncancerous cells, feature extraction, nuclei count, etc. The nuclei detection of a cell is an important task in the overall segmentation process. There are many existing techniques to localize the cell nuclei in 2D whole slide microscopic images. About 302 surgically excised Ki-67 labelled breast cancer specimens were subjected to Digital Image Analysis (DIA) on WSI samples. The tumour classifier method is used to identify tumour tissue automatically, but it hasn't been taught to differentiate between invasive and non-invasive cancer cells [98]. Cell nuclei yield the quantitative information to find the disease and its impact. In this regard, it is proposed an unsupervised machine learning for detecting cell nuclei and their segmentation using the matching-based technique. This method is validated on a total of 25 E&S liver histopathology images and 35 Papanicolaou-Stained thyroid images [99].

5.2 Materials and Methods

5.2.1 BreCaHAD Dataset

The BreCaHAD dataset of 162 breast cancer biopsy image samples prepared and released by the University of Calgary. It contains microscopic histopathological specimen at 40x magnification and keeps it in uncompressed image style .tiff contains a 3-color band of RGB with 8-bit depth in each band. The breast cell biopsy slides are stained with H&E. It is freely available and accessible on figshare at “https://doi.org/10.6084/m9.figshare.73791 86”. “The Alberta Health Research Ethics Board” has approved the relevant ethical approvals (HREBA.CC-17-0631). Various malignant cases are included in the dataset. This study gathered data from a variety of patients during their usual diagnoses. The University of Calgary was responsible for the preparedness and digitization of the dataset. Patients were not harmed in any way for the purpose of research [100].

5.2.2 Proposed Methodology to Grade Breast Cancer

All the important steps that deal with the identification of nuclear section, segmentation, pixel clustering on color intensity, and texture of nuclear sections are described in Figure 5.1. It is established that all the described techniques were executed under the regulated standards and guidelines. The decision of benign and malignant specimen tissue and its prognosis is dependent on the proliferation score (PS). The Ki-67 is the strong biomarker for calculating the proliferation score and describing the classification of the cells. The research work proposed an automatic detection of cell nuclei with the help of certain mathematical parameters to enhance the efficiency and reduce the duration of diagnosis spend by a pathologist to take a correct decision. The algorithm is coded in python programming language (Python 3.6, 32-bit) and it is supported by various open-source library packages like imutils, numpy, opencv, sklearn, matplotlib, and scikit-image.

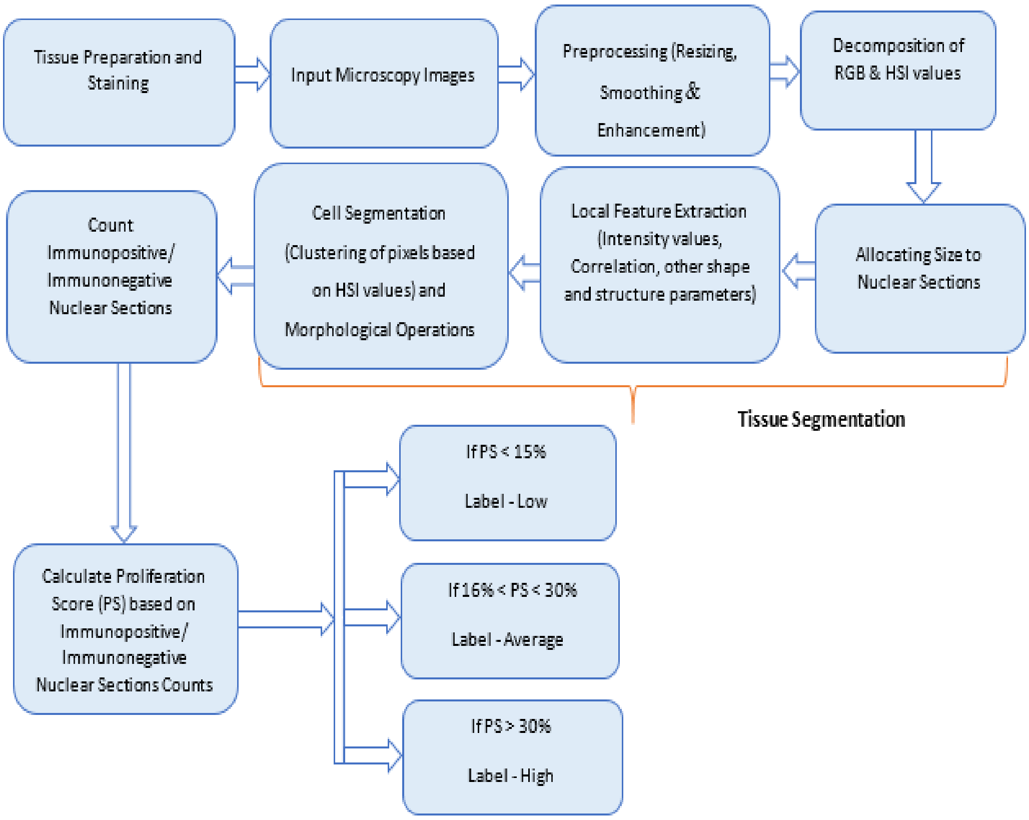

Tissue Preparation and Staining
Input Microscopy Images
Preprocessing (Resizing, Smoothing & Enhancement)
Decomposition of RGB & HSI values
Count Immunopositive/ Immunonegative Nuclear Sections
Cell Segmentation (Clustering of pixels based on HSI values) and Morphological Operations
Local Feature Extraction (Intensity values, Correlation, other shape and structure parameters)
Allocating Size to Nuclear Sections
Tissue Segmentation
If PS < 15%
Label - Low
Calculate Proliferation Score (PS) based on Immunopositive/ Immunonegative Nuclear Sections Counts
If 16% < PS < 30%
Label - Average
If PS > 30%
Label - High

Figure 5.1: Workflow of proposed unsupervised leveling of Ki-67 on BreCaHAD dataset images and quantifying the proliferation score

5.2.2.1 Tissue Preparation, Staining and Image Acquisition

The Breast Cancer Histopathology Annotation and Diagnosis (BreCaHAD) dataset contains 162 breast cancer biopsy pictures that can be used to evaluate and improve the proposed method's performance and efficiency [100][101]. Under controlled clinical conditions, the pathologist does a biopsy and execute various staining process. Hematoxylin and eosin (H&E) staining was employed on histological images annotated as mitosis, apoptosis, tumour nuclei, non-tumor nuclei, tubule, or non-tubule in this study. Each sample has a resolution of 1360 x 1024 pixels with a 0.514m x 0.527m per pixel at a magnification of x40 and is saved in uncompressed (.TIFF) format, 3-color band (RGB) with 8-bit depth in each band, and resolution of each sample is 1360 x 1024 pixels with a 0.514m x 0.527m per pixel at a magnification of x40. Each slide is manually focused [102].

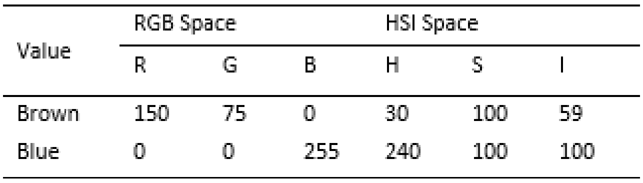
5.2.2.2 Preprocessing and Decomposition of Data Samples

It is discussed that the k-means clustering algorithm is unsupervised learning used to make the clusters of nuclei based on similarity. Before using the input images directly for feature extraction, it is always preferred to follow the steps of pre-processing. It results to improve the quality of input images. Usually, H&E stained images have problems like pigment gathering on stained tissues and abnormal distribution of small pigment particles around the tissue [94]. In this work, it is combined the Gaussian filter for smoothing and the laplacian filter of size 3x3 for enhancement of image quality. It is also used Otsu’s thresholding technique having different parameters that can help to identify the nuclei [103].

5.2.2.3 Tissue Segmentation

Tissue segmentation includes three consecutive steps as shown in Figure 5.1. After allocating the diameter of a nuclear section by pathologist selected on an input image, another local feature extraction is then implemented. It is suggested to include important parameters that help to extract local features of nuclei. The first parameter is the B channel (blue color intensity) in RGB and the second is the H channel (hue intensity) in HSI color space. Both are intrinsic characteristics of the H&E Ki-67 antigen. Brown and blue colour levels play an important role in distinguishing immunonegative from immunopositive nuclei [104]. It is shown in Table 5.1, the color intensity of brown is 0 in the blue band and 30 in the hue channels, while the same for blue is much higher 255 and 240 in RGB and HSI color space respectively. The difference between the two bands generates one of the highly relevant features for classification. Because of the characteristics of H&E, both the channels are highly scattered in different nuclei. To support the above points and color decomposition from original images is shown in Figure 5.2.

Table 5.1: Illustrating brown and blue color values in   RGB and HSI channels



Value	RGB Space			HSI Space		
	R	G	B	H	S	I
Brown	150	75	0	30	100	59
Blue	0	0	255	240	100	100

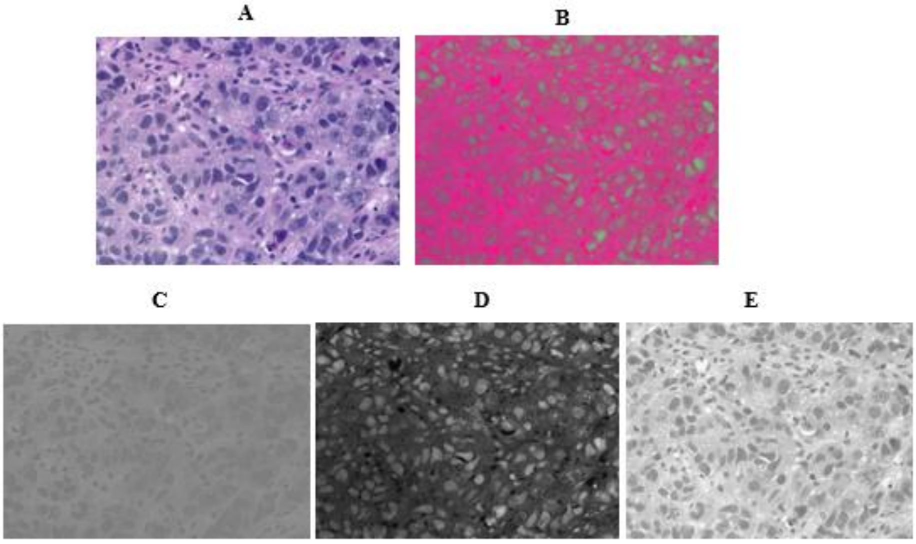


Figure 5.2: Representation of color decomposed from original images. (A) Sample of BreCaHAD dataset image, (B) HSI conversion of sample (A), and (C), (D), (E) corresponds to Hue, Saturation and Intensity channel respectively of HSI space

The mean and standard deviation of each pixel and behavior of its 3 x 3 neighborhood pixels are the second parameter. In the blue and hue intensity channels, the third parameter is local texture features, which include kurtosis and skewness of a single pixel and its 3 x 3 neighborhood pixels, respectively. The morphological operations like chess-board distance measurement and watershed boundaries algorithms can try to solve the problem of merged and superimposed nuclei. For solving such types of problems, it is suggested mathematical parameters related to formalizing the texture of different nuclei, which will help to improve the automation. It is discussed some important shape, structure, and morphological features that will help to identify the nuclear sections.

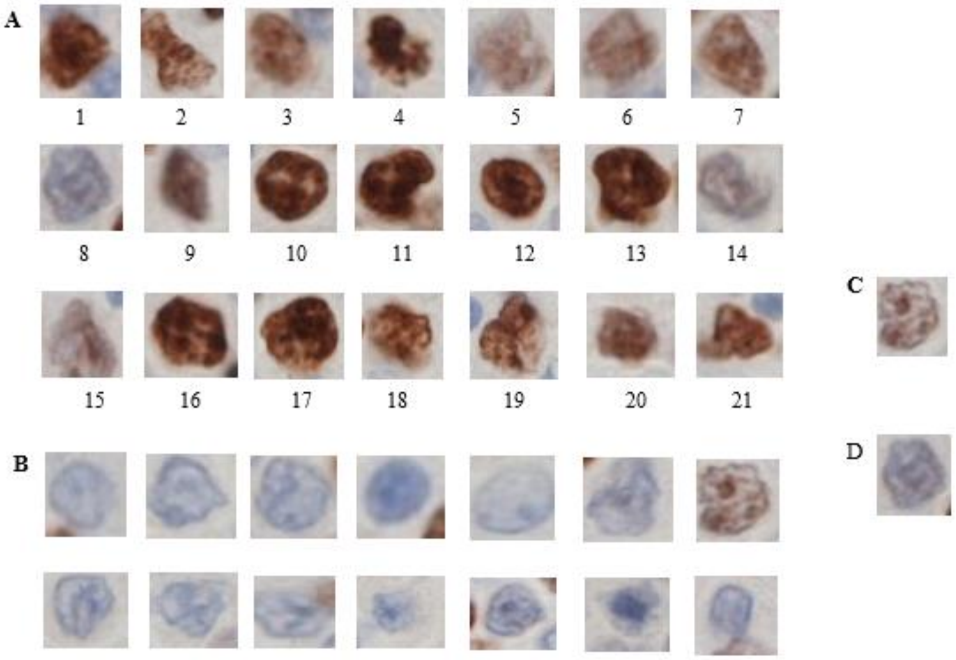
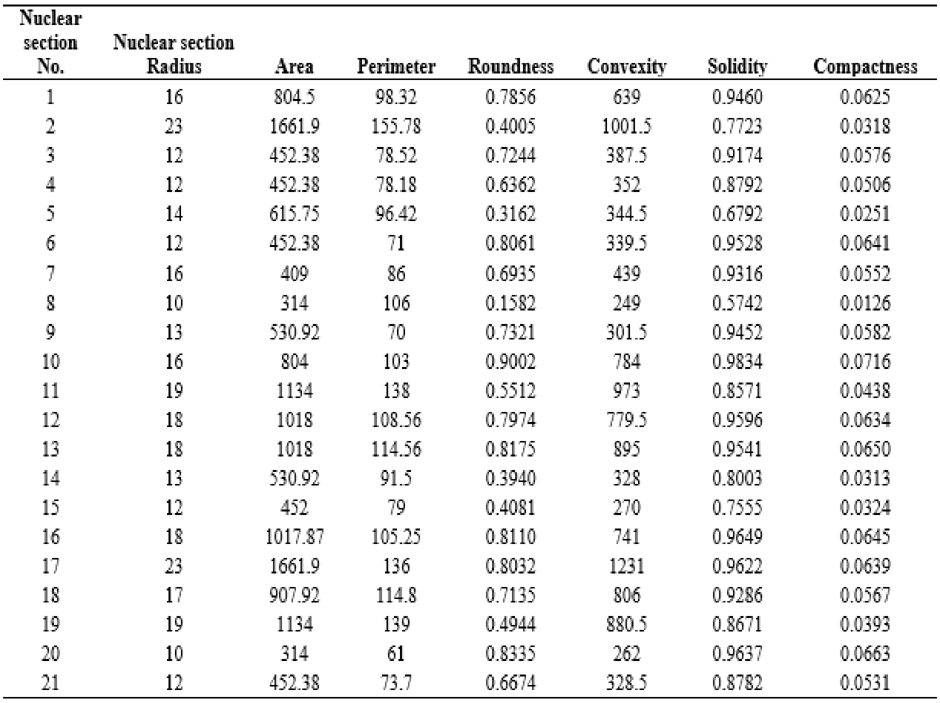


Figure 5.3: List of different texture and colors of Ki-67 nuclear sections (A) Sample of Immunopositive Nuclear Sections, (B) Sample of Immunonegative Nuclear Sections, (C) False negative nuclear section, and (D) False positive nuclear section

In Figure 5.3, it is shown that there are few samples of and immunopositive and immunonegative nuclear sections [105]. There are samples of false-negative and false-positive nuclear sections. It helps to develop a more effective and optimized decision-making algorithm. There is a substantial link between Ki-67 antigen levels and histology cancer grade. To find the correct counting of immunonegative and immunopositive nuclei, the morphological and biological parameter characteristics must be strong. It may be possible that some of the nuclear sections may be wrongly classified that can harm the final proliferation score and the error can propagate up to the final grading of cancer for a specific specimen sample. Hence, at some instant, the proposed algorithm can improve the grading result because of the cumulative effect of intensity, structure, and shape parameters on the nuclear section and can help to reduce the discrepancies. Figure 5.3, it is shown the different textures and colors of immunopositive nuclear sections.

Table 5.2: Illustrating morphological shape and feature values corresponding to Immunopositive Nuclear Sections in Figure 5.3(A)



Nuclear section No.	Nuclear section Radius	Area	Perimeter	Roundness	Convexity	Solidity	Compactness
1	16	804.5	98.32	0.7856	639	0.9460	0.0625
2	23	1661.9	155.78	0.4005	1001.5	0.7723	0.0318
3	12	452.38	78.52	0.7244	387.5	0.9174	0.0576
4	12	452.38	78.18	0.6362	352	0.8792	0.0506
5	14	615.75	96.42	0.3162	344.5	0.6792	0.0251
6	12	452.38	71	0.8061	339.5	0.9528	0.0641
7	16	409	86	0.6935	439	0.9316	0.0552
8	10	314	106	0.1582	249	0.5742	0.0126
9	13	530.92	70	0.7321	301.5	0.9452	0.0582
10	16	804	103	0.9002	784	0.9834	0.0716
11	19	1134	138	0.5512	973	0.8571	0.0438
12	18	1018	108.56	0.7974	779.5	0.9596	0.0634
13	18	1018	114.56	0.8175	895	0.9541	0.0650
14	13	530.92	91.5	0.3940	328	0.8003	0.0313
15	12	452	79	0.4081	270	0.7555	0.0324
16	18	1017.87	105.25	0.8110	741	0.9649	0.0645
17	23	1661.9	136	0.8032	1231	0.9622	0.0639
18	17	907.92	114.8	0.7135	806	0.9286	0.0567
19	19	1134	139	0.4944	880.5	0.8671	0.0393
20	10	314	61	0.8335	262	0.9637	0.0663
21	12	452.38	73.7	0.6674	328.5	0.8782	0.0531

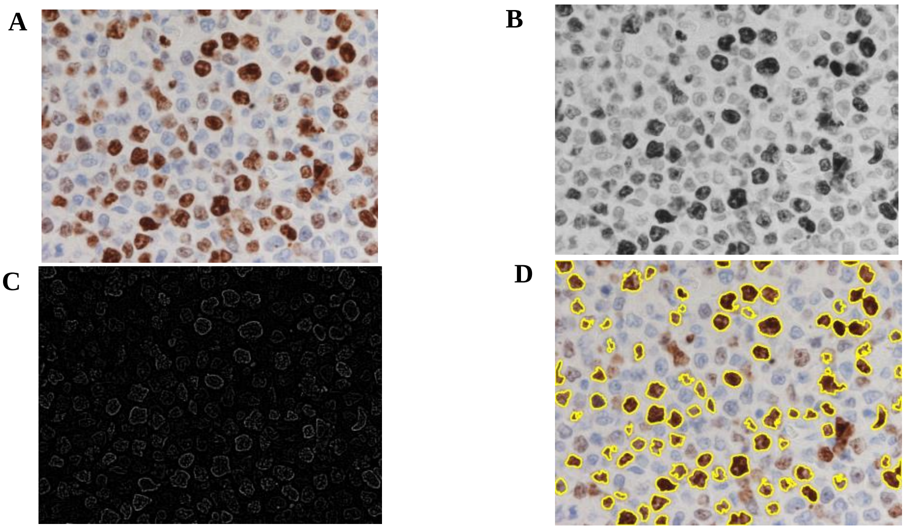


Figure 5.4: Results of immunopositive nuclear sections segmentation without clustering (A) Sample of 40x zoomed stained image, (B) Gray level image of input image using Otsu's threshold value, (C) Converted gray level image into contours, and (D) Automatic segmented immunopositive nuclear sections using various shape and structure features

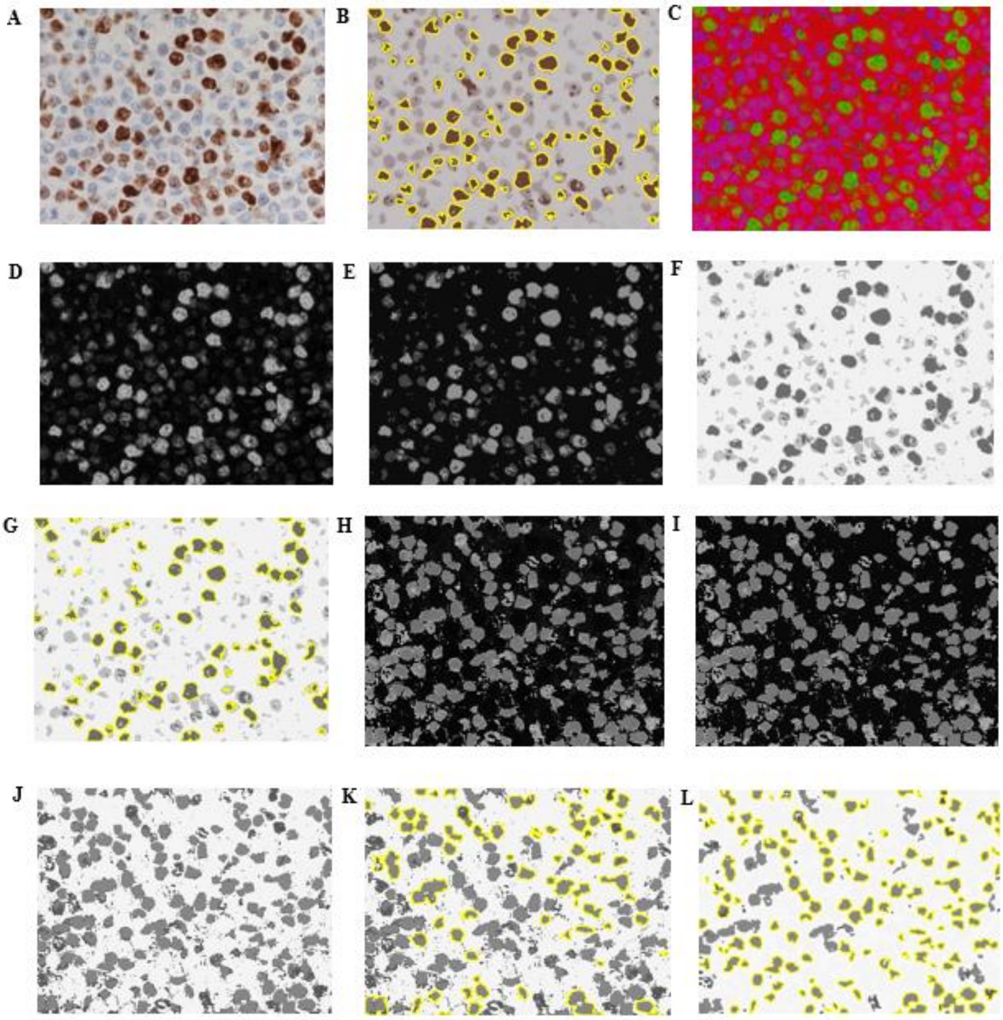


Figure 5.5: Results of immunopositive nuclear sections segmentation using k-means clustering (A) Sample of 40x zoomed stained image, (B) Output of sample (A) after clustering, and applying automatic segmented immunopositive nuclear sections using various shape and structure features mentioned in Table 4.1, (C) Representing HSI after conversion of original BGR image into HSI, (D) Representing saturation channel of HSI image, (E) Illustrating k-means clustering on saturation channel, (F) Invert the output of clustering, (G) Representing automated immunopositive nuclear sections, (H) Representing hue channel of HSI image, (I) Representing output of k-means clustering on hue channel, (J) Showing inversion of clustering, (K) Representing automated immunonegative nuclear sections and (L) Representing automated segmentation after applying morphological operators like Closing and Dilation to improve the accuracy of immunonegative nuclear sections counting

The sample images are morphologically and mathematically listed in Table 5.3. It contains the various important parameters that are sufficient to describe the size and structure of any nuclear section like its radius, area, convexity, solidity, perimeter, roundness, and compactness. Segmentation and automatic identification of immunopositive nuclei are demonstrated in Figure 5.4. To improve the segmentation process and automatic identification of immunopositive and immunonegative nuclei as demonstrated in Figure 5.4, it is adopted the k-means clustering algorithm, unsupervised learning that is implemented on HSI images as shown in Figure 5.5 [105].

5.2.2.4 Proliferation Score to Grade Breast cancer

After all the above calculations and getting the final count of immunopositive and immunonegative nuclear sections, calculate the proliferation score using Equation 5.1 [104][106].

$$Proliferation\ Score\ (PS)\ OR\ (ki-67\ Index) = \frac{Number\ of\ Immunopositive\ nuclear\ sections}{Total\ number\ of\ nuclear\ sections}\ \text{x}\ 100 \qquad (5.1)$$

Where,

$$Total\ number\ of\ nuclear\ sections = Number\ of\ immunopositive\ nuclear\ sections + Number\ of\ immunonegative\ nuclear\ sections$$

Based on the proposed methodology, the counting of immunopositive nuclear sections is 110 and immunonegative nuclear sections are 193 as illustrated in Figure 5.5. The proliferation score for the known counting after applying Equation 5.1 is calculated by 36.3%. After following the percentage of grading based on the proliferation score mentioned in Figure 5.1, it is found that the label of the sample stained image is **high**.

5.3 Results and Discussion

To the validation of the proposed method, it is using the total available 162 H&E stained BreCaHAD dataset which allows the researcher to optimize, evaluate, and usefulness of the proposed algorithm. The images have the dimension of 1360 x 1024 with a 40x magnifier in uncompressed (.TIFF) image format. The method is written in python programming language and deployed on PC with a configuration of 2GHz CPU and 4 GB RAM. It took 7.4 seconds on average to execute the single sample. If the configuration of the system will be improved, it will reduce the execution time and will adequate to meet the clinical real-time requirements. It is compared and confirmed the automated nuclear section segmentation against the manual nuclear section segmentation, and also evaluated the consistency after counting of nuclear sections for the proposed algorithm.

Another major challenge is the overlapping of the nuclear section that encourages to inaccurately count the number of nuclear sections. It has been observed that due to limitation of acquisition capability, it happens mostly the overlapping of the nuclear section. It is mentioned the problem of overlapping and it has found the solution with the watershed algorithm [37]. It is effectively working but it has also found some limitations. In Figure 5.5L, the paper has implemented morphological operators like opening and closing. It is helping to split the overlapped nuclear section as well as the unwanted region of interest and increases the counting accuracy. It is discussed more in the result section. In Figure 5.6, it is illustrating the problem of overlapping of nuclei.

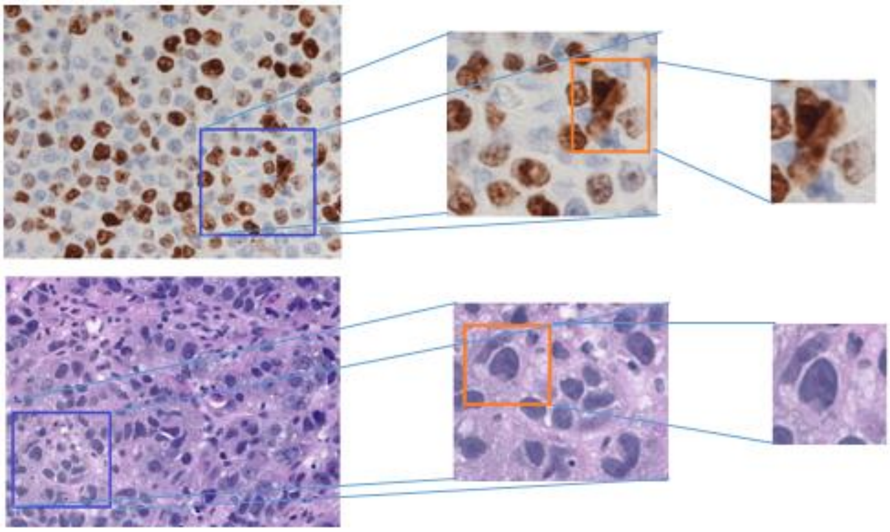


Figure 5.6: Sample of H&E stained images with the problem of overlapping of the nuclear section

Under the steps of tissue segmentation in Figure 5.1, it is mentioned to allocating size to the nuclear section. It is not always possible that the size of nuclear sections will be fixed for all samples as well as each level of magnification. As this method is assisting the pathologist for prediction purposes, so with the consent of the user (pathologist), will be selected two points on the sample image that will be strongly directed for the diameter of the nuclear section. This part of consent is not required for this study. It is known that counting is a major issue for grading cancer and overlapping of the nuclear section can affect the results. To minimize this, calculate the area of the ideal nuclear section as already mentioned by the user (pathologist). There will be a threshold band for the area of the region and other parameters like roundness, solidity, compactness, etc. If the area will be less than the lower threshold band, it will be considered as no nuclear section and it will not be counted. If the measured parameters are in between the lower and upper band it will be counted as one. Otherwise, if the area is above the upper band of the threshold value, generally it will be the case of overlapping and the algorithm will increase the number of counts accordingly.

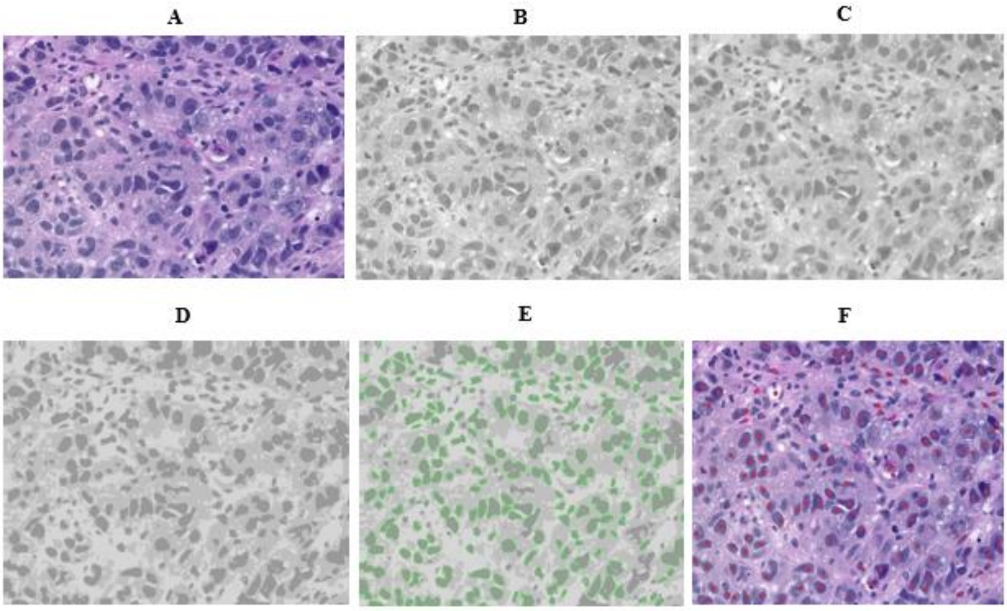


Figure 5.7: Accurate and efficient segmentation of BreCaHAD dataset image (A) Representing original BreCaHAD dataset, (B) Representing intensity channel of HSI image, (C) Result of dilation preceded by closing of intensity channel, (D) Result after applying k-means clustering, (E) More accurate automated segmentation result with good counting accuracy, and (F) Ground truth of sample (A) and the nuclear sections are in red circle

The effectiveness and performance of the above classifier are judged over the confusion matrix, for which is calculated the value of TP, TN, FP, and FN. The definition and formula are illustrated below:

Accuracy. Accuracy of any classification method defined as the counting of exactly classified samples i.e., TP and TN is shown as [107]:

$$Accuracy = \frac{TP+TN}{\mathrm{N}} * 100 \qquad 5.2$$

$Where,\ N = TP + TN + FP + FN$

Sensitivity. The sensitivity is calculated as the number of positive samples divided by the number of negative samples which are rightly classified [37] and is calculated as:

$$Sensitivity = \frac{TP}{FN+TP} \qquad 5.3$$

Where the sensitivity ranges varies from 0 to 1. '0' signifies worst and '1' best classification.

Specificity. The specificity is defined as the ratio of negative samples which are rightly classified [108]. It is shown as:

$$Specificity = \frac{TN}{FP+TN} \qquad 5.4$$

Where the specificity ranges varies from 0 to 1. '0' signifies worst and '1' best classification.

F-Measure. It is defined as the harmonic mean of recall and precision. It is shown as:

$$Precision = \frac{TP}{FP+TP}, \qquad Recall = \frac{TP}{FN+TP} \qquad 5.5$$

$$F - Measure = 2 * \frac{Precision * Recall}{Recall+Precision} \qquad 5.6$$

Where the value of F-measure ranges varies from 0 to 1. '0' signifies worst and '1' best classification.

Balanced Classification Rate (BCR). It is defined as the geometric mean of sensitivity and specificity is recognized as a balanced classification rate [9]. It is shown as:

$$BCR = \sqrt{Sensitivity * Specificity} \qquad 5.7$$

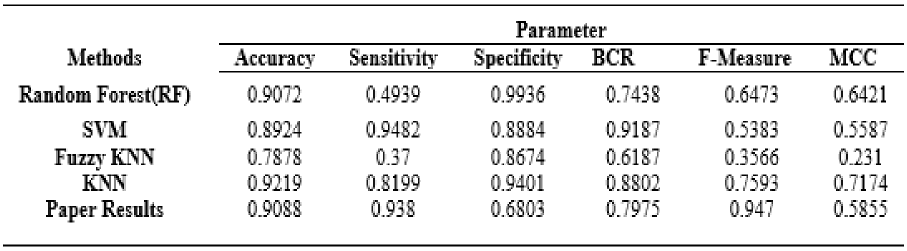
Mathew's Correlation Coefficient MCC). It is a measure of the eminence of binary classifications [108]. It is defined as:

$$MCC = \frac{TN*TP-FN*FP}{\sqrt{((FN+TP)(FP+TP)(FN+TN)(FP+TN))}} \qquad 5.8$$

The measure of MCC ranges from -1 to +1, where -1, +1 and 0, signifies the worst, best, and random predictions respectively.

Ethics and consent statements. The BreCaHAD dataset is freely available and accessible on Figshare at "https://doi.org/10.6084/m9.figshare.73791 86". All the necessary ethics approval has been granted by the "Health Research Ethics Board of Alberta (HREBA.CC-17-0631)" [100].

Table 5.3: Quantitative comparison of Different Method vs Paper Results based on proposed methodology



Methods	Parameter					
	Accuracy	Sensitivity	Specificity	BCR	F-Measure	MCC
Random Forest(RF)	0.9072	0.4939	0.9936	0.7438	0.6473	0.6421
SVM	0.8924	0.9482	0.8884	0.9187	0.5383	0.5587
Fuzzy KNN	0.7878	0.37	0.8674	0.6187	0.3566	0.231
KNN	0.9219	0.8199	0.9401	0.8802	0.7593	0.7174
Paper Results	0.9088	0.938	0.6803	0.7975	0.947	0.5855

Figure 5.7 is illustrated the output of all the phases of the proposed methodology. It is also adopted morphological operations like closing and opening to improve the segmentation results to better identify the nuclei. Table 5.4 is representing the performance comparison of the proposed model with other's work in same domain. It is observed that after applying this algorithm on 162 different BreCaHAD datasets, the value of F-score is recommendable, accuracy is approximate 0.9088, and sensitivity is 0.938 which also better than others, BCR is 0.7975 which is also better than Random Forest and Fuzzy KNN. It is illustrated in Figure 5.8, the H&E stained WSI sample images, the automated localization of nuclear sections where the nuclear sections are encircled by thin lines, and the manually annotated and labeled sample images by experienced pathologists in blue dots in Figure 5.8(A, B, and C) respectively.

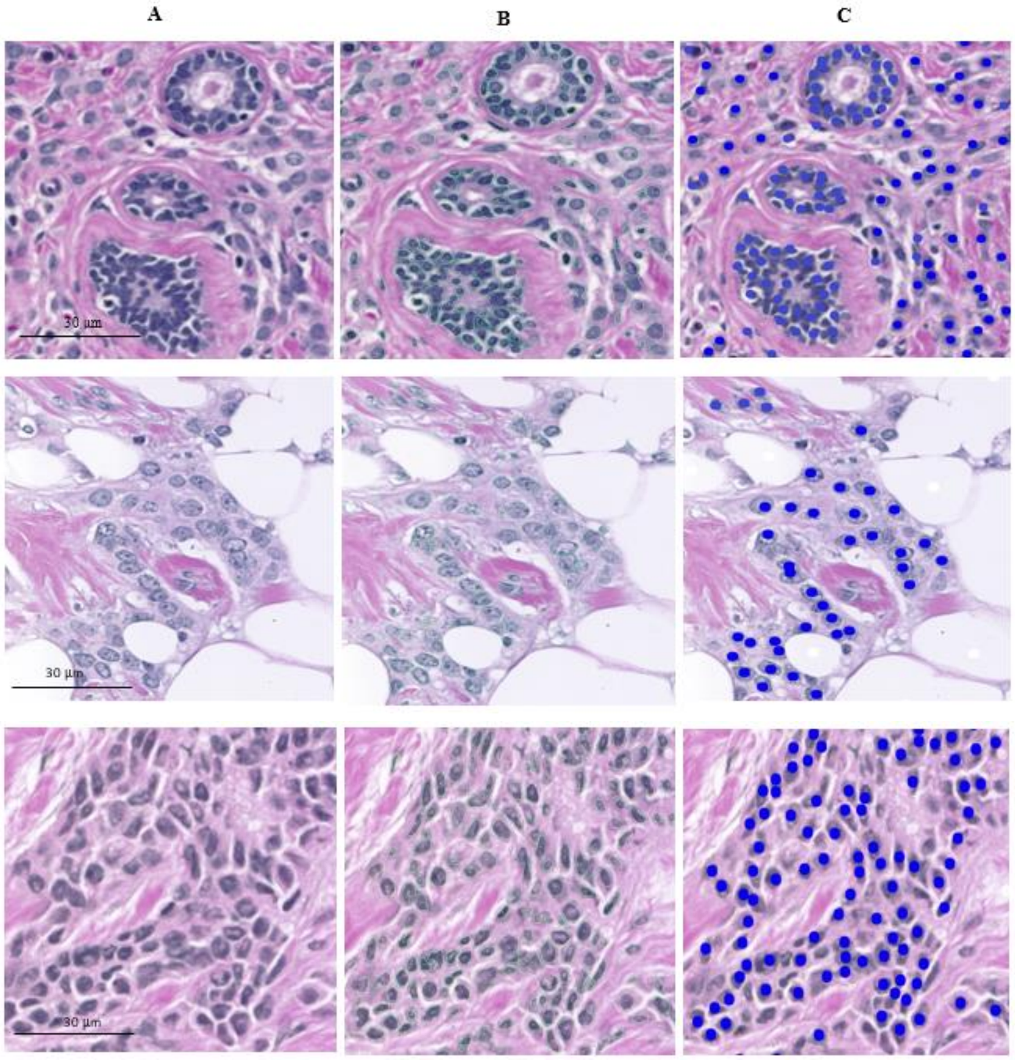


Figure 5.8: Localization of nuclear sections on H&E stained WSI image samples (a) H&E stained WSI samples of breast cancer tissues, (b) Automated localized nuclear sections, and (c) Samples labeled and annotated by medical experts of the BreCaHAD dataset

All the comparisons with a graphical representation of the segmentation algorithm with different performance parameters are shown in Figure 5.9. Potentially, this method will be highly supportive to the pathologists for fast, efficient, and accurate computation of Ki-67 proliferation score on breast H&E stained cancer images. If the number of sample images will increase, it will improve the accuracy and other dimensions of the model. The proposed method is compared to the different existing models as shown in Table 5.6.

Table 5.4: Comparison with Existing Methods vs Proposed Methodology

Research Papers	Comparison Parameters					
	Image Type	Sample Size	Image Size	Image Magnification	Methodology used	Accuracy (%) and other measures
Krishnan et al. (2011) [109]	Histology Images	158	800 x 800	40x	Support vector machine (SVM)	88.38
N. Khan et al. (2014) [110]	Neuroendocrine Tumor	57	10000 × 5000	40x	Conventional technique (Perceptual clustering)	94.60 (implemented on a very small dataset)
Kumar R., (2015) [37]	Histology Images	2828	Not Available	5x, 10x, 20x, and 40x	KNN	Sensitivity = 94.01, Specificity = 81.99, Recall = 64.6, Accuracy = 92.19, BCR =88.02, F-measure = 75.94, MCC = 71.74
P. Shi et al.(2016) [94]	Human nasopharyngeal carcinoma Xenografts	100	2040 × 1536	40x	Conventional techniques (smoothing, color channel decomposition, local feature extraction, K-means, watershed segmentation)	Mean Accuracy=75.1± 6.7%, where $\sigma_d = 6.7\%$
Paramanandan M., et al. (2016) [111]	Histopathology Images	39	1024 x 1280	40x	LBP algorithm on a MRF	90
Nawaz W., et al. (2018) [90]	H&E stained images	400	2048 x 1536	20x	CNN (ALEXNET)	81.25
Awan R, et al. (2018) [112]	Histology Images	400	512 x 512	20x	CNN+SVM	83.33
Aresta G., et al. (2019) [50]	Histology Images	500	2048 x 1536	40x	CNN	87
Proposed Methodology	Histology Images	162	1360 x 1024	40x	Machine Learning supported by morphological operators	90.8

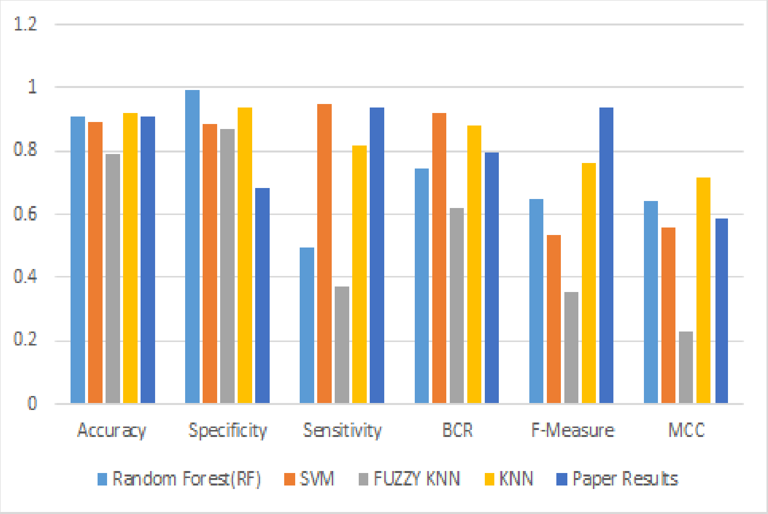


Figure 5.9: Graphical representation of a comparison of different segmentation algorithms like Random Forest, SVM, Fuzzy KNN, and KNN with performance parameters like accuracy, specificity, sensitivity, BCR, F-measure, and MCC

5.4 Chapter Summary

This chapter is mainly focusing on prognostic evaluation and grading of breast cancer. This section of work is implemented on the BreCaHAD dataset to automatic identification of nuclei used for tissue segmentation. The Ki-67 antigen is targeted in this section to calculate the proliferation score that is used to grade breast cancer as low, medium, and high.

Chapter 6

Conclusion and Future Direction

.

6.1 Summary and Main Contributions

This research uses the BACH dataset from the ICIAR 2018 Grand Challenge. The proposed methodology was successfully localized the region of interest with an accuracy of 85.5% on a total of 10 WSI annotated testing cancerous samples. It uses IoU to allow unsupervised machine learning with morphological features and shape formulae. The proposed study focuses on locating the region of interest in order to assist pathologists in making accurate and timely decisions about the amount of malignancy and subsequent treatment. This algorithm can be inbuilt with the existing CAD solutions as an option in the tool. It could be a friendly and good assistant for the pathologist that can reduce the effort and span of diagnosis. Through this novel research findings, one of the key obstacles of all neural network learning-driven systems, it has been discovered, is the availability of labelled data, which must be real. It is necessary to tune the neural network classification model for diverse datasets on a regular basis.

6.2 Contributions

The accuracy of localizing the ROI is found 85.5% and measured by similarity metrics intersection over union (IoU). It outperforms a variety of state-of-the-art algorithms. Pathologists use their specific knowledge to make diagnoses from sample photos. The pathogenic notion and computer vision features have a semantic mismatch. The proposed model is adaptive for such changes and mismatch after concerning to pathologists, it can be added the more cell features to make the method more robust and it will be more helpful to the pathologists to take a correct decision.

The BreCaHAD dataset was used to create an efficient automated nuclei segmentation model for cancer grading. The proposed work is based on using Ki-67 antigen that helps to differentiate between immunopositive and immunonegative cells. The achieved accuracy of the segmentation is 90.8% that is better than many existing algorithms.

6.3 Future Research Directions

For future perspective and scope of improvement, the work can be extended to various authentic datasets based on availability. If there are enough and diverse annotated WSI sample datasets, the outcome can be improved. Hardware advancements are, of course, as significant. This work is only implemented on H&E stained images. Other staining methods can be also implement on these samples where it is to be tested. Although, due to unavailability, it is not done. One of the most common problems of WSI sample images is the overlapping of the nuclei. It is very difficult to correct the counting of nuclei if the number of overlapping increases. Another limitation is the different biomarkers. The color and structural behavior of the tissues varies according to the different biomarkers. So it is one of the important points of concern. There is no standard benchmark for the annotation of the dataset. This work has the potential to commercialize if supported by the Cancer Research Center regarding the availability of WSI samples to increase its performance.

References

[1] World Cancer Report, “Cancer Research for Cancer Prevention,” 2020.

[2] GBD 2016 Risk Factors Collaborators, “Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016,” *Lancet.*, Vol. 390(10100), pp. 1345–1422, 2017.

[3] C. K. Zhou *et al.*, “Prostate cancer incidence in 43 populations worldwide: An analysis of time trends overall and by age group,” *International journal of cancer*, Vol. 138, No. 6, pp. 1388–1400, Mar. 2016.

[4] J. A. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J Clin.*, Vol. 68, No. 6, pp. 394–424, 2020.

[5] M. K. Mallath *et al.*, “The growing burden of cancer in India: epidemiology and social context.,” *The Lancet. Oncology*, Vol. 15, No. 6, pp. e205-12, May 2014.

[6] W. A. W. and K. S. Imran Ali, “Cancer Scenario in India with Future Perspectives,” *Cancer Therapy*, Vol. 8, pp. 56–70, 2011.

[7] “Three-Year Report of Population Based Cancer Registries (2012-2014), Bengaluru, India.”

[8] R. K. S. C. S. Pramesh, Rajendra A. Badwe, “The National Cancer Grid of India,” *Indian Journal of Medical and Paediatric Oncology*, Vol. 35, No. 3, 2014.

[9] J. Gerdes, U. Schwab, H. Lemke, and H. Stein, “Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation.,” *International journal of cancer*, Vol. 31, No. 1, pp. 13–20, Jan. 1983.

[10] L. Barisoni, K. J. Lafata, S. M. Hewitt, A. Madabhushi, and U. G. J. Balis, “Digital pathology and computational image analysis in nephropathology.,” *Nature reviews. Nephrology*, Vol. 16, No. 11, pp. 669–685, Nov. 2020.

[11] W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle, “Understanding and using DICOM, the data interchange standard for biomedical imaging,” *Journal of the American Medical Informatics Association : JAMIA*, Vol. 4, No. 3, pp. 199–212, 1997.

[12] Bruce H. McCormick, “Knife-Edge Scanning Microscope (KESM 1.5): Optics and Cameras T: tamu-cs-tr-2006-10-1,” 2006.

[13] H. A. Alturkistani, F. M. Tashkandi, and Z. M. Mohammedsaleh, “Histological Stains: A Literature Review and Case Study.,” *Global journal of health science*, Vol. 8, No. 3, pp. 72–79, Jun. 2015.

[14] T. Imamura, T. Saitou, and R. Kawakami, “In vivo optical imaging of cancer cell function and tumor microenvironment.,” *Cancer science*, Vol. 109, No. 4, pp. 912–918, Apr. 2018.

[15] A. Milbourne *et al.*, “Results of a pilot study of multispectral digital colposcopy for the in vivo detection of cervical intraepithelial neoplasia,” *Gynecologic oncology*, Vol. 99, No. 3 Suppl 1, pp. S67-75, 2005.

[16]

“www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm552742.htm.” .

[17] B. P. Riordan DP, Varma S, West RB, “Automated Analysis and Classification of Histological Tissue Features by Multi-Dimensional Microscopic Molecular Profiling.,” *PLoS ONE*, Vol. 10, No. 7, 2015.

[18] H. S. Mousavi, V. Monga, G. Rao, and A. U. K. Rao, “Automated discrimination of lower and higher grade gliomas based on histopathological image analysis,” *Journal of pathology informatics*, Vol. 6, p. 15, Mar. 2015.

[19] N. N. Cirean DC, Giusti A, Gambardella LM, Schmidhuber J. In: Mori

K, Sakuma I, Sato Y, Barillot C, "Mitosis Detection in Breast Cancer Histology Images with Deep," in *Medical image computing and computer-assisted intervention MICCAI 2013:16th international conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013*, 2013.

[20] T. J. Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, "Potential Roles for Spectroscopic Coherent Raman Imaging for Histopathology and Biomedicine," in *4th IEEE International Symposium*, 2007, p. 1284e7.

[21] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 284–287, 2008.

[22] K. M. Meiburger *et al.*, "Validation of the Carotid Intima-Media Thickness Variability: Can Manual Segmentations Be Trusted as Ground Truth?," *Ultrasound in medicine & biology*, Vol. 42, No. 7, pp. 1598–1611, Jul. 2016.

[23] J. M. EH. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt and Ogden, "Pyramid Methods in Image Processing," *RCA Engineer*, Vol. 29, No. 6, 1984.

[24] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of Tumor Histology via Morphometric Context," *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2013, p. 10.1109/CVPR.2013.286, Jun. 2013.

[25] S. H. Heywang-Köbrunner, A. Hacker, and S. Sedlacek, "Advantages and Disadvantages of Mammography Screening.," *Breast care (Basel, Switzerland)*, Vol. 6, No. 3, pp. 199–207, 2011.

[26] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep Learning to Improve Breast Cancer Detection on Screening Mammography.," *Scientific reports*, Vol. 9, No. 1, p. 12495, Aug. 2019.

[27] S. P. Rana *et al.*, “Machine Learning Approaches for Automated Lesion Detection in Microwave Breast Imaging Clinical Data.,” *Scientific reports*, Vol. 9, No. 1, p. 10510, Jul. 2019.

[28] B. Sahiner *et al.*, “Malignant and benign breast masses on 3D US volumetric images: effect of computer-aided diagnosis on radiologist accuracy.,” *Radiology*, Vol. 242, No. 3, pp. 716–724, Mar. 2007.

[29] Z. Klimonda, P. Karwat, K. Dobruch-Sobczak, H. Piotrzkowska-Wróblewska, and J. Litniewski, “Breast-lesions characterization using Quantitative Ultrasound features of peritumoral tissue.,” *Scientific reports*, Vol. 9, No. 1, p. 7963, May 2019.

[30] C. Meeuwis *et al.*, “Computer-aided detection (CAD) for breast MRI: evaluation of efficacy at 3.0 T.,” *European radiology*, Vol. 20, No. 3, pp. 522–528, Mar. 2010.

[31] B. S. Ko *et al.*, “MRI-based 3D-printed surgical guides for breast cancer patients who received neoadjuvant chemotherapy.,” *Scientific reports*, Vol. 9, No. 1, p. 11991, Aug. 2019.

[32] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, “Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model,” *Scientific Reports*, Vol. 7, No. 1, p. 4172, 2017.

[33] H. Fox, “Is H&E morphology coming to an end?,” *Journal of clinical pathology*, Vol. 53, No. 1, pp. 38–40, Jan. 2000.

[34] 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.

[35] M. Can, A., Bello, M.O., Cline, H.E., Tao, X., Ginty, F., Sood, A., Gerdes, M.J., & Montalto, “Multi-modal imaging of histological tissue sections,” in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2008, pp. 288–291.

[36] S. Liao, M. W. K. Law, and A. C. S. Chung, “Dominant Local Binary Patterns for Texture Classification,” *IEEE Transactions on Image*

Processing, Vol. 18, No. 5, pp. 1107–1118, 2009.

[37] R. Kumar, R. Srivastava, and S. Srivastava, "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features," *Journal of Medical Engineering*, Vol. 2015, p. 457906, 2015.

[38] M. M. Zainudin, M. N. S., Mohd Said, M., & Ismail, "Feature Extraction on Medical Image Using 2D Gabor Filter," *Applied Mechanics and Materials, 52–54, 2128–2132. https://doi.org/10.4028/www.scientific.net/amm.52-54.2128*, Vol. 52–54, pp. 2128–2132, 2011.

[39] A. Kumar and M. Prateek, "Advancements in Cancer Diagnosis Using Digital Imaging System: A Review," *IJAST*, Vol. 29, No. 05, pp. 11242–11254, 2020.

[40] P. J. Han J, Kamber M, *Data Mining Concepts and Techniques*, 3rd ed. 225Wyman Street,Waltham, MA 02451, USA: Morgan Kaufmann Publishers is an imprint of Elsevier, 2012.

[41] H. S. Kanghee Park, Amna Ali, Dokyoon Kim, Yeolwoo An, Minkoo Kim, "Robust predictive model for evaluating breast cancer survivability," *Engineering Applications of Artificial Intelligence*, Vol. 26, pp. 2194–2205, 2013.

[42] O. A. Filipczuk P., Kowal M., "Automatic Breast Cancer Diagnosis Based on K-Means Clustering and Adaptive Thresholding Hybrid Segmentation," *In: Choraś R.S. (eds) Image Processing and Communications Challenges 3. Advances in Intelligent and Soft Computing Springer, Berlin, Heidelberg.*, Vol. 102, 2011.

[43] C. M. A, "color clustering technique for image segmentation," *Computer Vision, Graphics and Image Processing*, Vol. 52, No. 2, pp. 145–170, 1990.

[44] J. H. S. Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, "Patch-based Convolutional Neural Network for Whole Slide

Tissue Image Classification," in *Computer Vision and Pattern Recognition*, 2016.

[45] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images.," *Journal of the American Medical Informatics Association : JAMIA*, Vol. 20, No. 6, pp. 1099–1108, 2013.

[46] J. D. Webster and R. W. Dunstan, "Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology.," *Veterinary pathology*, Vol. 51, No. 1, pp. 211–223, Jan. 2014.

[47] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, Vol. 7, p. 29, Jul. 2016.

[48] J.-M. Chen *et al.*, "New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images.," *Scientific reports*, Vol. 5, p. 10690, May 2015.

[49] D. Tellez *et al.*, "Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks.," *IEEE transactions on medical imaging*, Mar. 2018.

[50] G. Aresta *et al.*, "BACH: Grand challenge on breast cancer histology images.," *Medical image analysis*, Vol. 56, pp. 122–139, Aug. 2019.

[51] J. Yanase and E. Triantaphyllou, "The seven key challenges for the future of computer-aided diagnosis in medicine," *International journal of medical informatics*, Vol. 129, pp. 413–422, 2019.

[52] J. Melamed *et al.*, "The cooperative prostate cancer tissue resource: a specimen and data resource for cancer researchers.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, Vol. 10, No. 14, pp. 4614–4621, Jul. 2004.

[53] M. Uhlen *et al.*, "Towards a knowledge-based Human Protein Atlas.,"

Nature biotechnology, Vol. 28, No. 12. United States, pp. 1248–1250, Dec-2010.

[54] "Comprehensive genomic characterization defines human glioblastoma genes and core pathways.," *Nature*, Vol. 455, No. 7216, pp. 1061–1068, Oct. 2008.

[55] E. C. Shaw *et al.*, "Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study.," *Journal of clinical pathology*, Vol. 65, No. 5, pp. 403–408, May 2012.

[56] M. Lundin, J. Lundin, and J. Isola, "Virtual microscopy.," *Journal of clinical pathology,* Vol. 57, No. 12. pp. 1250–1251, Dec-2004.

[57] M. N. Lassere, "The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate end," *Statistical methods in medical research*, Vol. 17, No. 3, pp. 303–340, Jun. 2008.

[58] C. M. Micheel, S. J. Nass, and G. S. Omenn, Eds., *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington (DC), 2012.

[59] R. Colomer *et al.*, "Biomarkers in breast cancer: A consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology.," *Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, Vol. 20, No. 7, pp. 815–826, Jul. 2018.

[60] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond.," *Annual review of pathology,* Vol. 8, pp. 331–359, Jan. 2013.

[61] T. Kohlberger *et al.*, "Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection," *Journal of pathology informatics*, Vol. 10, p. 39, Dec. 2019.

[62] S. Kothari *et al.*, “Automatic batch-invariant color segmentation of histological cancer images,” *Proceedings. IEEE International Symposium on Biomedical Imaging*, Vol. 2011, pp. 657–660, 2011.

[63] Z. He and W. Yu, “Stable feature selection for biomarker discovery.,” *Computational biology and chemistry*, Vol. 34, No. 4, pp. 215–225, Aug. 2010.

[64] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, “Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer,” *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, Vol. 2011, pp. 422–425, Nov. 2011.

[65] D. Zhang and G. Lu, “Review of shape representation and description techniques,” *Pattern Recognition*, Vol. 37, No. 1, pp. 1–19, 2004.

[66] J. C. Caicedo, F. A. González, and E. Romero, “Content-based histopathology image retrieval using a kernel-based semantic annotation framework.,” *Journal of biomedical informatics*, Vol. 44, No. 4, pp. 519–528, Aug. 2011.

[67] E. A. Abdel-Zaher AM, “Breast cancer classification using deep beliefnetworks,” *Expert Syst Appl*, Vol. 46, pp. 139–44, 2016.

[68] W. Sun, T.-L. B. Tseng, J. Zhang, and W. Qian, “Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data.,” *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, Vol. 57, pp. 4–9, Apr. 2017.

[69] T. Sadad, A. Munir, T. Saba, and A. Hussain, “Fuzzy C-means and region growing based classification of tumor from mammograms using hybrid texture feature,” *J. Comput. Sci.*, Vol. 29, pp. 34–45, 2018.

[70] B. Mughal, N. Muhammad, M. Sharif, A. Rehman, and T. Saba, “Removal of pectoral muscle based on topographic map and shape-shifting silhouette.,” *BMC cancer*, Vol. 18, No. 1, p. 778, Aug. 2018.

[71] B. Mughal, M. Sharif, N. Muhammad, and T. Saba, “A novel

classification scheme to decline the mortality rate among women due to breast tumor," *Microsc ResTech*, Vol. 81, pp. 171–80, 2017.

[72] B. Mughal, N. Muhammad, M. Sharif, T. Saba, and A. Rehman, "Extraction of breast border and removal of pectoral muscle in wavelet domain," *Biomedical Research-tokyo*, Vol. 28, pp. 5041–5043, 2017.

[73] A. A. Duarte MA, Pereira WC, "Calculating texture features frommammograms and evaluating their performance in classifying clusters of microcalcifications," in *Mediterranean Conference on Medical and BiologicalEngineering and Computing*, 2019, pp. 322–32.

[74] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform," *Measurement*, Vol. 146, pp. 800–805, 2019.

[75] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunyé, and J. G. Elmore, "Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images: a Comparative Study.," *Journal of digital imaging*, Vol. 29, No. 4, pp. 496–506, Aug. 2016.

[76] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges.," *Journal of infection and public health*, Vol. 13, No. 9, pp. 1274–1289, Sep. 2020.

[77] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, "Biological Interpretation of Morphological Patterns in Histopathological Whole-Slide Images.," *ACM-BCB ... ...: the ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, Vol. 2012, pp. 218–225, Oct. 2012.

[78] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2017.," *CA: a cancer journal for clinicians*, Vol. 67, No. 1, pp. 7–30, Jan. 2017.

[79] R. A. Smith, V. Cokkinides, and H. J. Eyre, "American Cancer Society

Guidelines for the Early Detection of Cancer, 2005.," *CA: a cancer journal for clinicians*, Vol. 55, No. 1, pp. 31–36, 2005.

[80] J. R. Gilbertson, J. Ho, L. Anthony, D. M. Jukic, Y. Yagi, and A. V Parwani, "Primary histologic diagnosis using automated whole slide imaging: a validation study.," *BMC clinical pathology*, Vol. 6, p. 4, Apr. 2006.

[81] "ICIAR 2018: Grand challenge on breast cancer histology images, https://iciar2018-challenge.grandchallenge.org, (accessed 11 March 2018)." .

[82] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62–66, 1979.

[83] W. R. Gonzalez RC, *Digital Image Processing*, 3rd ed. Prentice Hall, 2002.

[84] A. Kumar and M. Prateek, "Localization of Nuclei in Breast Cancer Using Whole Slide Imaging System Supported by Morphological Features and Shape Formulas.," *Cancer management and research*, Vol. 12, pp. 4573–4583, 2020.

[85] S. S. Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression," in *Computer Vision and Pattern Recognition*, 2019.

[86] G. Apou, B. Naegel, G. Forestier, F. Feuerhake, and C. Wemmert, "Efficient Region-based Classification for Whole Slide Images BT - Computer Vision, Imaging and Computer Graphics - Theory and Applications," 2015, pp. 239–256.

[87] Z. Zhang *et al.*, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Machine Intelligence*, Vol. 1, No. 5, pp. 236–245, 2019.

[88] Z. Guo *et al.*, "Publisher Correction: A Fast and Refined Cancer Regions

Segmentation Framework in Whole-slide Breast Pathological Images," *Scientific Reports*, Vol. 10, No. 1, p. 8591, 2020.

[89] J. Z. Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, "Ensemble Network for Region Identification in Breast Histopathology Slides," in *In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, 2018.

[90] K. H. A. Nawaz W., Ahmed S., Tahir A., "Classification Of Breast Cancer Histology Images Using ALEXNET," in *In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, Springer, Cham, 2018, pp. 869–876.

[91] G. Aditya, D. Anand, and S. Amit, "Classification of Breast Cancer Histology Using Deep Learning," in *In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, Springer, Cham, 2018, pp. 837–844.

[92] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, "Patch-based system for Classification of Breast Histology images using deep learning.," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, Vol. 71, pp. 90–103, Jan. 2019.

[93] R. Yan *et al.*, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, Vol. 173, pp. 52–60, 2020.

[94] P. Shi, J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen, "Automated Ki-67 Quantification of Immunohistochemical Staining Image of Human Nasopharyngeal Carcinoma Xenografts.," *Scientific reports*, Vol. 6, p. 32127, Aug. 2016.

[95] F. Penault-Llorca and N. Radosevic-Robin, "Ki67 assessment in breast cancer: an update.," *Pathology*, Vol. 49, No. 2, pp. 166–171, Feb. 2017.

[96] C. M. Perou *et al.*, "Molecular portraits of human breast tumours,"

Nature, Vol. 406, No. 6797, pp. 747–752, 2000.

[97] R. Yerushalmi, R. Woods, P. M. Ravdin, M. M. Hayes, and K. A. Gelmon, “Ki67 in breast cancer: prognostic and predictive potential.,” *The Lancet. Oncology*, Vol. 11, No. 2, pp. 174–183, Feb. 2010.

[98] L. Harris *et al.*, “American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer,” *Journal of Clinical Oncology*, Vol. 25, No. 33, pp. 5287–5312, Nov. 2007.

[99] R. Nishimura, T. Osako, Y. Okumura, M. Hayashi, and N. Arima, “Clinical significance of Ki-67 in neoadjuvant chemotherapy for primary breast cancer as a predictor for chemosensitivity and for prognosis.,” *Breast cancer (Tokyo, Japan),* Vol. 17, No. 4, pp. 269–275, Oct. 2010.

[100] D. A, Aksac, DJ and A. T, Özyer, R, “BreCaHAD: A Dataset for Breast Cancer Histopathological Annotation and Diagnosis,” *BMC Research Notes,* 2018.

[101] A. Kumar and M. Prateek, “Automated Detection and Classification of Ki-67 Stained Nuclear Section Using Machine Learning Based on Texture of Nucleus to Measure Proliferation Score for Prognostic Evaluation of Breast Carcinoma,” *Preprint, Research Square,* 2020.

[102] C. Liu, F. Shang, J. A. Ozolek, and G. K. Rohde, “Detecting and segmenting cell nuclei in two-dimensional microscopy images,” *Journal of pathology informatics*, Vol. 7, p. 42, Oct. 2016.

[103] P. W. Van Vliet, L. J., Young, L. T. & Verbeek, “Recursive gaussian derivative filters,” in *Fourteenth International Conference on Pattern Recognition*, 1998, pp. 509–514.

[104] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, “An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer,” *Scientific Reports*, Vol. 7, No. 1, p. 3213, 2017.

[105] https://www.memorangapp.com/flashcards/60092/1.01+-

+Cell+Cycle+Control/, “Cite for Figure: Ki-67 stained image.” .

[106] J. Cottenden *et al.*, “Validation of a Cytotechnologist Manual Counting Service for the Ki67 Index in Neuroendocrine Tumors of the Pancreas and Gastrointestinal Tract.,” *Archives of pathology & laboratory medicine,* Vol. 142, No. 3, pp. 402–407, Mar. 2018.

[107] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 24, No. 7, pp. 971–987, 2002.

[108] L. Wei, Y. Yang, and R. M. Nishikawa, “Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis,” *Pattern recognition,* Vol. 42, No. 6, pp. 1126–1132, Jun. 2009.

[109] C. Muthu Rama Krishnan, M.; Chakraborty, R. R. Paul, and A. Ray K, “Quantitative Analysis of Sub-Epithelial Connective Tissue Cell Population of Oral Submucous Fibrosis Using Support Vector Machine,” *Journal of Medical Imaging and Health Informatics,* Vol. 1, No. 1, pp. 4–12, 2011.

[110] M. K. Khan Niazi, M. M. Yearsley, X. Zhou, W. L. Frankel, and M. N. Gurcan, “Perceptual clustering for automatic hotspot detection from Ki-67-stained neuroendocrine tumour images.,” *Journal of microscopy,* Vol. 256, No. 3, pp. 213–225, Dec. 2014.

[111] M. Paramanandam *et al.*, “Automated Segmentation of Nuclei in Breast Cancer Histopathology Images.,” *PloS one,* Vol. 11, No. 9, p. e0162053, 2016.

[112] N. R. Ruqayya Awan, Navid Alemi Koohbanani, Muhammad Shaban, Anna Lisowska, “Context-Aware Learning using Transferable Features for Classification of Breast Cancer Histology Images,” in *Computer Vision and Pattern Recognition (cs.CV),* 2018.

Publications

1. A. Kumar and M. Prateek, “Advancements in Cancer Diagnosis Using Digital Imaging System: A Review,” *IJAST*, Vol. 29, No. 05, pp. 11242–11254, 2020.
2. A. Kumar and M. Prateek, “Localization of Nuclei in Breast Cancer Using Whole Slide Imaging System Supported by Morphological Features and Shape Formulas.,” *Cancer management and research*, Vol. 12, pp. 4573–4583, 2020.
3. A. Kumar and M. Prateek, “Automated Detection and Classification of Ki-67 Stained Nuclear Section Using Machine Learning Based on Texture of Nucleus to Measure Proliferation Score for Prognostic Evaluation of Breast Carcinoma,” *Preprint, Research Square*, 2020. (Under Preprint)