

**DEVELOPMENT OF DEEP LEARNING BASED  
AUTOMATED METHODS FOR BREAST CANCER  
HISTOPATHOLOGY IMAGE ANALYSIS**

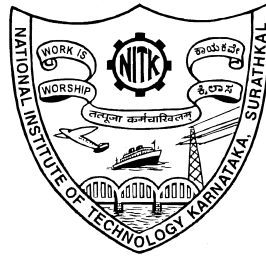
**THESIS**

Submitted in partial fulfillment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

by

**TOJO MATHEW**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA  
SURATHKAL, MANGALORE - 575025, INDIA**

**JUNE 2022**



## DECLARATION

I hereby *declare* that the Research Thesis entitled **Development of Deep Learning based Automated Methods for Breast Cancer Histopathology Image Analysis** which is being submitted to the *National Institute of Technology Karnataka, Surathkal* in partial fulfillment of the requirements for the award of the Degree of *Doctor of Philosophy* is a *bona fide report of the research work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.



**TOJO MATHEW**

Registration No.: 165005CS16P01

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Surathkal - 575025

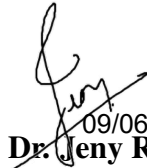
Place: NITK - Surathkal

Date: 09-JUNE-2022



## CERTIFICATE

This is to *certify* that the Research Thesis entitled **Development of Deep Learning based Automated Methods for Breast Cancer Histopathology Image Analysis**, submitted by **TOJO MATHEW** (Registration No: 165005CS16P01) as the record of the research work carried out by him, is *accepted* as the *Research Thesis submission* in partial fulfillment of the requirements for the award of degree of *Doctor of Philosophy*.



09/06/2022  
**Dr. Jeny Rajan**

Research Guide

Assistant Professor

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Surathkal-575025



09-06-2022

Dr. Shashidhar G. Koolagudi

**Chairman - DRPC**

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Surathkal-575025

(Signature with Date and Seal)



*To my wife Susan, children  
Aaron & Sharron..*





## ACKNOWLEDGEMENTS

We are at the end of the line! Even though it felt far away at first, I now know how quickly time travels. Fortunately, I had a great time on this voyage, and I was able to make a lot of new friends, travel around, cooperate with outstanding colleagues, and acquire a lot of vital life lessons. This achievement would not have been possible without the support of my numerous friends, co-workers, and my family. I'd like to offer my heartfelt gratitude to each and every one of you.

I have been fortunate to have my research guide, Dr. Jeny Rajan, Assistant Professor, Department of CSE, NITK Surathkal, India, who gave me enormous freedom to explore on my own, and well-timed guidance when my steps faltered. His meticulous and concise comments and thoughtful criticism was instrumental in making my endeavor materialized. His endless support, trust, patience and honest feedback made this achievement possible. Thank you sir.

I would also like to take this opportunity to convey my heartfelt reverence to my research collaborator and mentor, Dr. Jyoti R. Kini, Professor, Department of Pathology, Kasturba Medical College, Mangalore, India for her invaluable guidance and encouragement throughout my research.

I am greatly indebted to the highly insightful experts in my Doctoral Research Progress Assessment Committee Dr. Basavaraj Talwar, Assistant Professor, Department of CSE, NITK Surathkal, India and Dr. A.V. Narasimhadhan, Assistant Professor, Department of ECE, NITK Surathkal, India. Their timely suggestions and the continual flow of ideas through the constructive feedback leads to the fulfilment of my research thesis.

I wholeheartedly express my gratitude to Ms. Vani M, Dr. P. Santhi Thilagam, Dr. Alwyn Roshan Pais, Dr. Shashidhar G Koolagudi, Heads of the Department (during my period of study), Department of Computer Science, NITK Surathkal. I stand obliged before them for the affection and kindness bestowed towards me. I feel immensely proud to acknowledge the facilities and kindhearted support by Prof. Udaykumar R. Yaragatti, respected Director of NITK Surathkal.

I am grateful to the management committee & the leadership of 'The National Institute of Engineering' (NIE, Mysuru) for the support and facilities provided to me during these years.

A special heartfelt appreciation and thanks to my wife Susan Daniel, and our most beloved children Aaron & Sharron for all the unconditional support, care and love which

they have provided and their immense sacrifices throughout my research period.

Always the most important part, my deep gratitude goes to my father (Late Mr. Mathew), mother (Ms. Alice), brothers (Binu Mathew, Anumod Mathew), father-in-law (Mr. M Daniel), mother-in-law (Ms. Saroja Rani), cousins (Joseph Joshi, Jiji Alexander), other cousins, uncles, aunts and grandparents, for their unconditional love, support and encouragement in my life.

With great gratitude, I remember all my teachers who have taught me at various stages of my education. Their blessings and support have enabled me to come so far in life. Especially, I remember Mr. V.V Madhavan and Dr. Wilscy M who have been my inspiration and role models in pursuing my teaching and research dreams.

I am lucky to have many nice and kind friends around me. I would like to thank them all, especially Dr. Johnpaul CI & family, Dr. Krishnakumar, Yamanappa, Dr. Anoop, Niyas, Pawan, Ajith B, Dr. Amit, Sachin, Bijay, Akhila, Dr. Girish, Sudhish, Ajnas, Dr. Nikhil, Dr. Pramod, Dr. Christina, Dr. Swathi for being supportive and helpful in many difficult situations.

I would love to thank everyone who is directly and indirectly contributed to the successful completion of my doctoral research work.

Finally, but most importantly, I thank the Almighty who accompanies me throughout this journey in the form of all the people mentioned above and others, situations, challenges, opportunities, inner voices, and whatnots..., disseminating little pieces of wisdom and reasons to move on every time. It has been an awesome journey so far, hand-in-hand with the Creator, through the pathways of life to my destiny.

TOJO MATHEW

Place: NITK - Surathkal

Date: 09-JUNE-2022

## ABSTRACT

According to the recent report by the Global Cancer Observatory, breast cancer has overtaken lung cancer as the leading type of cancer in terms of new cases reported. In 2020, breast cancer accounted for 11.7% of all new cancer cases and 6.9% all cancer related deaths. Timely diagnosis and targeted treatment can significantly improve the survival chances of breast cancer patients. Pathological procedures are integral parts of cancer diagnosis and treatment planning. In the routine cancer pathology analysis, tissue samples are extracted from the tumor regions and applied with suitable staining agents. The glass slides prepared this way are analyzed by pathologists through a microscope to make interpretations about the disease condition. The manual procedure of microscopy analysis is tedious, time consuming, and error-prone. Digitization of pathological glass slides into slide images opens a plethora of possibilities to apply computational methods to automate several pathology procedures. The focus of this thesis work is to develop computational methods for automated analysis of breast cancer histopathology images and extract clinically relevant information to support prognosis and treatment planning. Grading and molecular subtyping of breast cancer are the two important pathology procedures considered for automation in this thesis work. Particularly, automation of two breast cancer grading procedures namely *mitosis detection* and *nuclear atypia scoring* are taken as the first two objectives. The third objective is automated *molecular subtyping* of breast cancer, a classification that supports targeted treatment and hence better outcome.

*Breast cancer grading* categorizes the disease based on its aggressiveness. The grade information is used for prognosis and treatment planning. Among the three parameters involved in breast cancer grading (mitotic cell count, nuclear atypia score, and

tubule formation), mitotic cell counting is the most challenging task for pathologists. It is possible to automate this task by applying computational algorithms on pathology slide images. Lack of sufficiently large datasets, and class imbalance between mitotic and non-mitotic cells are the two major challenges in developing effective deep learning-based methods for automated *mitosis detection*. In order to address these challenges, an approach of combining datasets from different sources and a more effective image data augmentation technique are used. Following these, a novel method pipeline is proposed which makes use of an advanced deep learning algorithm to address this problem. In contrast to the existing methods that are trained and validated on independent datasets, the proposed approach aims to develop generalized dataset-agnostic solutions for mitosis detection. The results obtained for the proposed method show improvement over existing deep learning methods based on independent datasets.

*Nuclear atypia score* is the second parameter used for grading breast cancer. Manual procedure of nuclear atypia scoring is laborious and marked by pathologists' disagreement as well as low reproducibility. Automation of this procedure using computational methods is seen as a viable alternative to these challenges. It is observed that most of the existing methods rely on extracted feature-based learning algorithms. Deep learning algorithms are not sufficiently utilized to address this task. In this thesis, a novel deep learning based framework for automated nuclear atypia scoring of breast cancer is proposed. The framework consists of three major phases namely preprocessing, deep learning, and postprocessing. In the proposed approach, the original three-class problem of slide level atypia scoring is reformulated as a six-class problem of nuclei classification for the effective use of deep learning algorithms. The method based on this framework gives a performance that exceeds the state-of-the-art by a significant margin.

*Molecular subtyping* classifies cancer based on the expression of genetic alterations behind the disease. Identifying the specific subtype aids in targeted treatment of the disease to achieve better outcome. Molecular subtyping through immunohistochem-

istry (IHC) analysis is a pathology procedure to determine the subtype of breast cancer. The existing manual procedure involves assessing the status of the four molecular biomarkers ER, PR, HER2, and Ki67. To automate this procedure, a deep learning-based framework using IHC image analysis is proposed. At present, there are no methods found in literature for IHC based automated molecular subtyping. The proposed system is evaluated for the performance of individual biomarker status predictions and patient-level subtype classification. The results obtained at the various levels of evaluations are highly promising.

In the extensive literature study carried in the preliminary stage of the research work, it is understood that the potential of deep learning algorithms is not fully utilized in the automation of pathology procedures for mitosis detection and nuclear atypia scoring. The bottlenecks for this are identified and potential solutions are investigated in this thesis work. The performance of proposed methods for these tasks validates the relevance of the solution approach adopted. In the absence of any prior work in the literature for automated molecular subtyping of breast cancer, the proposed deep learning-based classification framework establishes a new direction for automating this labor-intensive pathology procedure. The high performance of the proposed method is a strong indication of the clinical applicability of automated methods. In essence, by automating three key pathology procedures in breast cancer diagnosis and treatment planning, this thesis work aims to contribute to the global research efforts towards making cancer treatment more effective, affordable, and accessible.

*Keywords:* Histopathology; Breast Cancer; Cancer grading; Mitosis, Nuclear atypia; Deep learning; Convolutional neural networks; Patch extraction; Nuclei segmentation; Data augmentation; Immunohistochemistry; Molecular subtyping; Biomarkers; Estrogen receptor; Progesterone receptor; Ki67; Human epidermal growth factor receptor



# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>4</b>
<b>ABSTRACT</b>	<b>6</b>
<b>TABLE OF CONTENTS</b>	<b>13</b>
<b>LIST OF TABLES</b>	<b>15</b>
<b>LIST OF FIGURES</b>	<b>19</b>
<b>ABBREVIATIONS AND NOMENCLATURE</b>	<b>20</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Overview of Cancer and Histopathology . . . . .	1
1.2 Breast Cancer Histopathology . . . . .	3
1.2.1 Breast Cancer Grading . . . . .	3
1.2.2 Molecular Subtyping of Breast Cancer . . . . .	4
1.2.3 Automation of Pathology Procedures . . . . .	6
1.3 Motivation and Problem Statement . . . . .	8
1.3.1 Problem Statement . . . . .	9
1.4 Major Contributions . . . . .	11
1.5 Organization of this Thesis . . . . .	12
<b>2 LITERATURE SURVEY</b>	<b>13</b>
2.1 Breast Cancer Grading & Related Challenges . . . . .	14
2.2 Automated Mitosis Detection . . . . .	16
2.2.1 Methods using Handcrafted Features . . . . .	16

2.2.2	Methods using Deep Learning . . . . .	19
2.2.3	Combination Methods . . . . .	21
2.3	Automated Nuclear Atypia Scoring . . . . .	22
2.4	Automated Analysis of Molecular Biomarkers . . . . .	25
2.4.1	ER and PR . . . . .	25
2.4.2	HER2 . . . . .	27
2.4.3	Ki67 . . . . .	29
2.5	Summary . . . . .	31
<b>3</b>	<b>AUTOMATED MITOSIS DETECTION IN HISTOPATHOLOGY IMAGES</b>	<b>35</b>
3.1	Introduction . . . . .	35
3.1.1	Challenges in Automated Mitosis Detection . . . . .	36
3.2	Methodology . . . . .	38
3.2.1	Image Color Normalization . . . . .	39
3.2.2	Candidate Cell Detection & Segmentation . . . . .	40
3.2.3	Context Preserving Data Augmentation . . . . .	43
3.2.4	CNN Training and Evaluation . . . . .	44
3.3	Experimental Results & Discussion . . . . .	46
3.3.1	Dataset . . . . .	46
3.3.2	Experiment Setup . . . . .	47
3.3.3	Evaluation Metrics . . . . .	48
3.3.4	Results . . . . .	49
3.3.5	Discussion . . . . .	55
3.3.5.1	Impact of Combining Datasets . . . . .	55
3.3.5.2	Impact of Context Preserving Augmentation . . . . .	56
3.4	Summary . . . . .	57
<b>4</b>	<b>AUTOMATED NUCLEAR ATYPIA SCORING OF BREAST CANCER</b>	<b>59</b>



4.1	Introduction . . . . .	59
4.1.1	Challenges in Automated Nuclear Atypia Scoring . . . . .	60
4.1.1.1	Complexity of Histological Slide Images . . . . .	61
4.1.1.2	Inter-class Similarity and Intra-class Variations . . . . .	61
4.2	Proposed Framework . . . . .	65
4.2.1	Image Preprocessing . . . . .	67
4.2.2	Deep Learning Stage . . . . .	70
4.2.3	Postprocessing and Atypia Scoring . . . . .	74
4.3	Experimental Results & Discussions . . . . .	76
4.3.1	Dataset & Experimental Setup . . . . .	78
4.3.2	Results and Analysis . . . . .	80
4.3.2.1	Selection of Patch Size for CNN . . . . .	81
4.3.2.2	Performance of the CNN Classifier . . . . .	82
4.3.2.3	Evaluation and Comparison of Nuclear Atypia Scoring . . . . .	83
4.3.3	Discussion . . . . .	87
4.3.3.1	Pathologists' Disagreement and Labeling Discrepancies . . . . .	87
4.3.3.2	Nuclear Atypia Scoring: Man vs. Machine . . . . .	89
4.3.3.3	Future Prospects for the Proposed Framework . . . . .	90
4.4	Summary . . . . .	91
<b>5</b>	<b>AUTOMATED MOLECULAR SUBTYPING OF BREAST CANCER</b>	<b>93</b>
5.1	Introduction . . . . .	93
5.2	Methodology . . . . .	96
5.2.1	Training Pipeline . . . . .	96
5.2.1.1	Image Patch Extraction & Augmentation . . . . .	97
5.2.1.2	Training of CNN Models . . . . .	98
5.2.2	Evaluation Pipeline . . . . .	99

5.2.2.1	Pre-processing of IHC Images . . . . .	99
5.2.2.2	Image-patch Classification . . . . .	100
5.2.2.3	Post-processing for Biomarker Status Prediction . . . . .	102
5.2.2.4	Molecular Subtyping . . . . .	104
5.3	Experimental Results & Discussion . . . . .	104
5.3.1	Dataset and Experimental Setup . . . . .	105
5.3.1.1	Evaluation Metrics . . . . .	106
5.3.2	Results and Discussion . . . . .	106
5.3.2.1	Results of Image-patch Classification . . . . .	108
5.3.2.2	Results of Slide Level Biomarker Status Prediction . . . . .	110
5.3.2.3	Patient Level Biomarker Status Prediction . . . . .	114
5.3.2.4	Molecular Subtype Classification . . . . .	115
5.3.2.5	Discussions and Future Scope . . . . .	115
5.4	Summary . . . . .	117
<b>6</b>	<b>CONCLUSIONS</b>	<b>119</b>
	<b>APPENDICES</b>	<b>123</b>
<b>A</b>	<b>Evaluation Metrics</b>	<b>123</b>
<b>B</b>	<b>Methodology Adoption Pattern Over the Years for Automated Mitosis Detection</b>	<b>125</b>
<b>C</b>	<b>Adoption Pattern of Standard Algorithms for Automated Mitosis Detection</b>	<b>127</b>
	<b>REFERENCES</b>	<b>129</b>
	<b>LIST OF PAPERS BASED ON THESIS</b>	<b>145</b>
	<b>BIODATA</b>	<b>147</b>

## LIST OF TABLES

1.1	Nottingham Grading System (NGS) parameters and scoring criteria for breast cancer grading. . . . .	4
2.1	Summary of the handcrafted feature based mitosis detection methods presented in this study. . . . .	17
2.2	Summary of the deep learning-based mitosis detection methods. . . . .	20
2.3	Combination methods using handcrafted features & deep learning. . . . .	21
2.4	Summary of the major works in literature for automated atypia scoring . . . . .	24
2.5	Summary of the works related to IHC image analysis to assess the different cancer biomarkers . . . . .	30
3.1	Nuclei detection rate (mitotic) given by the detection algorithm (Al-Kofahi <i>et al.</i> , 2009) used in the proposed method. . . . .	42
3.2	Result of the experiments carried out to choose the suitable CNN architecture for the proposed method. . . . .	49
3.3	Result obtained using five-fold cross validation of base dataset MITOS-ATYPIA with CPDA. . . . .	50
3.4	Result obtained using 5-fold cross validation of combined dataset with CPDA. . . . .	50
3.5	Comparison of context preserving data augmentation (CPDA) and conventional context non-preserving data augmentation (CNDA) in each fold of 5-fold cross validation. . . . .	53
3.6	Comparison of the proposed method with the state-of-the-art deep learning methods based on MITOS-ATYPIA dataset. (A: Aperio scanner images, H: Hamamatsu scanner images, M: MITOS dataset). . . . .	54
4.1	Performance comparison of different deep CNNs considered for the six-class classifier in the proposed framework. . . . .	73
4.2	Performance of the nuclear image patch classifier model on different subsets of the MITOS-ATYPIA dataset. . . . .	82

4.3	Results of the proposed method and comparison with the state-of-the-art methods using Aperio scanner images. . . . .	84
4.4	Results of the proposed method and comparison with the state-of-the-art methods using Hamamatsu scanner images. . . . .	84
4.5	Results of the proposed method and comparison with the state-of-the-art methods using combined Aperio and Hamamatsu scanner images. . . . .	85
4.6	Comparison of average area under ROC curve (AUC) with state-of-the-art for different scanner image sets. . . . .	87
4.7	Degree of disagreement between the two pathologists in the first level scoring of the MITOS-ATYPIA Aperio scanner images. The subsets A10 and A14 show no disagreement whereas A18 shows scoring disagreement between the two pathologists for 50% of the images. . . . .	88
4.8	Comparison of the proposed method with the initial scoring of the MITOS-DATASET by two independent pathologists. . . . .	89
5.1	Molecular subtypes of breast cancer and their characteristics (Table created from Eliyatkin <i>et al.</i> (2015)) . . . . .	94
5.2	Parameters used for DenseNet CNN . . . . .	99
5.3	Details of the training image patches used to train CNN models used in the proposed method . . . . .	106
5.4	Result of the biomarker image patch classification by the DenseNet CNN used in the proposed method. . . . .	107
5.5	Result of the slide image classification to target biomarker status classes. . . . .	111
5.6	Result of applying different threshold criteria for Ki67 slide level status prediction as Ki67 High/Low. (PPP : Positive patch percentage, PPC: Positive patch count) . . . . .	112
5.7	Result of patient level biomarker status prediction. . . . .	113
5.8	Result of patient-wise molecular subtype classification. . . . .	115
A.1	Definition of confusion matrix elements used in various evaluation metrics. . . . .	123

## LIST OF FIGURES

1.1	A sample high power field (HPF) slide image captured at 40× magnification of the microscope. . . . .	6
1.2	Immunohistochemistry (IHC) slide image samples of biomarkers used for molecular subtyping of breast cancer, (a) Estrogen receptor (ER) , (b) Progesterone receptor (PR), (c) Human epidermal factor receptor 2 (HER2), (d) Antigen Ki67. . . . .	8
2.1	Organization of the literature survey . . . . .	13
2.2	Typical workflow of methods using handcrafted features. . . . .	16
2.3	Typical workflow of methods using deep learning. . . . .	19
3.1	Challenges in mitosis detection related to appearance of cells. (a) Shape and size variations among mitotic cells, (b) Non-mitotic cells or structures that resemble mitotic cells in appearance. . . . .	37
3.2	A graphical outline of the proposed method that involves a training pipeline and a testing pipeline. In the training pipeline, context preserving data augmentation (CPDA) is applied for the mitotic cell patches. The output of the training pipeline is a trained CNN model employed in the testing pipeline to classify the cell images. . . . .	39
3.3	Sample outputs of the Reinhard color normalization of H & E images. Images (a, b, c, & d) are original images from two different scanners of the MITOS-ATYPIA dataset (ICPR, 2014) and (e, f, g, & h) are the corresponding color-normalized images. . . . .	40
3.4	Candidate cell segmentation of HPF images. Images a, b, c are the normalized HPF images and d, e, f are the corresponding segmentation output. . . . .	41
3.5	(a) Representation of conventional context non-preserving data augmentation (CNDA), (b) Context preserving data augmentation (CPDA) used in the proposed method. . . . .	43
3.6	Learning pattern of the CNN using (a) base dataset and (b) combined dataset. . . . .	51

3.7	Precision-Recall curve and average precision (AP) obtained for the mitotic and non-mitotic cell classification using different folds of the combined dataset (a) Fold 1, (b) Fold 2, (c) Fold 3, (d) Fold 4, and (e) Fold 5. . . . .	52
3.8	Representative confusion matrices obtained for the combination of datasets MITOS and MITOS-ATYPIA (a) Fold 1, (b) All folds combined. Mitosis is the positive class and non-mitosis figures constitute the negative class in this binary classification problem. . .	52
4.1	Sample slide images to demonstrate complexity and structural diversity within histopathology slide images. Major portions of the slide images are occupied by (a) stroma, tumor cells, and lymphocytes, (b) tumor cells, necrosis, stroma, and fat globules. A closer view of different regions is given in Figure 4.8. . . . .	60
4.2	Inter-class similarity of slide images. From subset A10: (a) Score 2 slide image, (b) Score 3 slide image; From subset A11: (c) Score 2 slide image, (d) Score 3 slide image. . . . .	62
4.3	Intra-class variations in score 2 type slide images from different subsets (A03, A04 etc.) of the MITOS-ATYPIA dataset. These samples vary substantially in appearance even though they all have the same atypia score of 2. . . . .	63
4.4	Overview of the proposed framework for automated nuclear atypia scoring. Three major phases in the framework are: (a) Image preprocessing, (b) Deep learning, (c) Post-processing. (Post-processing is involved only in the slide level evaluation of nuclear atypia score.) .	66
4.5	Color normalization of slide images. (a, b, and c) are the unnormalized images and (d, e, and f) are the corresponding normalized images. .	68
4.6	Nuclei seed detection in slide regions of different nuclei classes. (a, b, c, and d) are the slide regions of lymphocytes, score 2, score 3, and necrosis classes respectively, (e, f, g, and h) are corresponding nuclei detected. . . . .	69
4.7	Training specific preprocessing and CNN training in the deep learning phase of the proposed framework. . . . .	70
4.8	Samples of class-wise region crops from the slide images for the six-class classifier model designed. Scoring classes (SC) of nuclear atypia: (a) Score 1, (b) Score 2, (c) Score 3; Elimination classes (EC): (d) Lymphocytes, (e) Stroma, (f) Necrosis. . . . .	72

4.9	Slide level of evaluation of nuclear atypia score. Input to the evaluation pipeline is the preprocessed test instances consisting of color-normalized slide images and nuclei seeds detected. Output is the nuclear atypia score of the input slide image. . . . .	75
4.10	Instances of plurality voting-based prediction of nuclear atypia score from Aperio and Hamamatsu scanner image sets. (GT: Ground truth score). . . . .	77
4.11	Result of experiments carried out to decide the input patch size to be used for the CNN. . . . .	81
4.12	Learning pattern of the six-class CNN classifier (DenseNet) used in the proposed method for images sets from (a) Aperio scanner, (b) Hamamatsu scanner. . . . .	82
4.13	ROC curve and AuC for the three scoring classes of tumor cells for image set (a) Aperio, (b) Hamamatsu, (c) Combined MITOS-ATYPIA dataset. . . . .	86
5.1	Training pipeline of the proposed framework for molecular subtyping. a) ER, PR, & Ki67 images are used to train three binary CNN classifier models separately for each biomarker. b) For HER2, a three-class CNN model is trained to classify each region patch into one of the output classes. . . . .	97
5.2	Evaluation pipeline of the proposed framework for automated molecular subtyping. . . . .	100
5.3	Extraction of foreground regions from the biomarker images to facilitate image patch extraction based on objects/regions of interest only: (a) Original IHC images of ER, PR, Ki67 & HER2; (b) Masks of detected foreground (white)/background (black); (c) Overlay of masks over the images. . . . .	101
5.4	Graphs of patch level classifications for Fold 1. Row 1: ER, Row 2: PR, Row 3: Ki67, Row 4: HER2. . . . .	109
5.5	Determination of optimal value for minimum positive patch count (cnt) per slide in Ki67 status prediction. Precision (pr) peaked when cnt is kept as 24, Recall (re) peaked for cnt value 18, and F1 score (fs) showed maximum value for cnt value 20. . . . .	112
B.1	Methodology adoption trend over the years . . . . .	125

C.1 Usage pattern of standard algorithms in the methods reviewed.  
(Acronyms: Convolutional neural network (CNN), Support Vector  
Machine (SVM), Random forest (RF), Decision tree (DT), Completed  
local binary pattern (CLBP), Maximum-likelihood estimation (MLE),  
Laplacian of Gaussian (LOG), Active contour model (ACM)) . . . . 127



## ABBREVIATIONS AND NOMENCLATURE

AP	Average Precision
AUC	Area under ROC curve
CNDA	Context Non-preserving Data Augmentation
CNN	Convolutional Neural Network
CPDA	Context Preserving Data Augmentation
EC	Elimination Classes
ER	Estrogen Receptor
FPR	False Positive Rate
H & E	Hematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
HPF	High Power Field
IDC	Invasive Ductal Carcinoma
IHC	Immunohistochemistry
LoG	Laplacian of Gaussian
NGS	Nottingham Grading System
PR	Progesterone Receptor
ROC	Receiver Operating Characteristic
SC	Scoring Classes
TMA	Tissue Microarray
TPR	True Positive Rate



# CHAPTER 1

## INTRODUCTION

This chapter provides an overview of the pathology procedures for grading and molecular subtyping of breast cancer, automation of these procedures through image analysis and its significance in improving cancer treatment. Motivation of the research, objectives and the major contributions are presented in the final sections of the chapter.

### 1.1 Overview of Cancer and Histopathology

Cancer refers to a group of diseases that are characterized by uncontrollable proliferation of cells in the living body that leads to the formation of tumors. It can affect any part of the human body such as lungs, liver, colon, stomach, breast, etc., and spread from the primary affected site to other parts of the body. Cancer costs millions of lives across the world every year. A recent report by the World Health Organization (WHO, 2020) says cancer caused around 10 million deaths across the world in the year 2020. The latest GLOBOCAN report (Sung *et al.*, 2020) estimated 19.3 million new cancer cases in 2020. The threat of cancer is projected to worsen with 28.4 million new cases in 2040. Early diagnosis and timely treatment can increase the survival chances of cancer patients to a large extent. Hence there are active research works ongoing worldwide to improve early diagnosis of cancer (Wardle *et al.*, 2015) and devise appropriate treatment. Considering the severe menace posed by cancer on humanity, there is a dire need to accelerate the research on cancer to improve the current treatment protocols.

Histopathology is a tissue level study of diseases for diagnostic and prognostic evaluation. It has a vital role in treatment of cancer. In histopathology analysis of cancer, the

tissues are extracted from the suspected tumor region through a biopsy procedure. Histology glass-slides are prepared using these tissues by following the routine procedures and analyzed via an appropriate microscope. Two important pathology procedures in the treatment of cancer are *grading* and *molecular subtyping*. Grading of cancer classifies the disease into different histologic grades by considering the factors like how different the appearance of the tumor cells/tissues is compared to the normal ones, and the tumor growth rate. Cancer grading primarily aims at determining the aggressiveness of cancer. Aggressiveness indicates how fast the tumor is growing and how likely it can spread to other parts of the body. These details help in improved prognosis and treatment planning of the disease. Cancer grading protocols vary for different cancer types. For example, prostate cancer grading is done using Gleason scoring (Epstein *et al.*, 2016), Fuhrman system (Fuhrman *et al.*, 1982) is used for renal carcinoma, and Anneroth/Bryne invasive front grading (Bryne *et al.*, 1989; Sawair *et al.*, 2003) for oral squamous cell carcinoma.

Another therapeutically relevant classification of cancer is the molecular subtyping (Collisson *et al.*, 2019; Al-Thoubaity, 2020; Guinney *et al.*, 2015). It considers the genetic factors behind malignancy to categorize the cancer into different subtypes and facilitates targeted therapy of the disease. Although gene expression profiling is the direct way to identify the genetic alterations that trigger malignancy, the procedure is costly and not routinely available. An alternate pathology procedure is to investigate the presence of tissue level bio-markers produced by the underlying genetic factors. All these pathology procedures are labor-intensive, time-consuming, and error-prone due to the human factors like fatigue, expertise etc. Moreover, manual procedures are known to have high levels of interobserver disagreement (Robbins *et al.*, 1995; Malon *et al.*, 2012).

## 1.2 Breast Cancer Histopathology

Breast cancer is a heterogeneous type of cancer that originates in breast tissues, and it primarily affects women. In 2020, female breast cancer has overtaken lung cancer to become the most dominant type of cancer globally. There were 2.26 million new breast cancer cases and 6,84,996 deaths in 2020 according to the GLOBOCAN report. These statistics portray an alarming picture. Breast cancer-related mortality can be reduced by early-stage detection and accurate identification of the specific subtype of the cancer (Harbeck and Gnant, 2016) to provide targeted treatment in a timely manner. Grading and molecular subtyping of breast cancer are significant procedures to achieve these objectives.

### 1.2.1 Breast Cancer Grading

Breast cancer is graded by a system known as the Nottingham Grading System (NGS) (Elston and Ellis, 2002). It is a modification of the Bloom-Richardson grading system (Bloom and Richardson, 1957). There are three parameters used for the grading of breast cancer as per the NGS. They are: *i) mitotic count*, *ii) nuclear pleomorphism (atypia)*, *iii) tubule formation*. Mitosis is the process of cell division in living organisms. The number of dividing cells in the tumor region is indicative of the growth rate of the tumor. Atypia scoring (Das *et al.*, 2020b) quantifies the size and the shape variations of cancer cells. Tubule formation (Basavanhally *et al.*, 2011) refers to the ring-like structures formed by the cancer cells and typically found in low grade cancers.

Among the three parameters in NGS, mitotic cell count is the most objective one. The other two parameters are relatively subjective in nature, and the scoring accuracy of these depends largely on the expertise of the pathologist. Each of these parameters is assigned a score ranging from 1 to 3 based on the criteria defined in NGS. The parameters of NGS and the scoring criteria for a specific configuration of the microscope is shown in Table 1.1. Based on the total score (TS) obtained by adding the individual

Table 1.1: Nottingham Grading System (NGS) parameters and scoring criteria for breast cancer grading.

Parameter	Score	Score Criteria
Mitosis count	1	0–9 mitotic cells in 10 consecutive High Power Fields (HPFs)
	2	10–19 mitotic cells in 10 consecutive HPFs
	3	$\geq 20$ mitotic cells in 10 consecutive HPFs
Nuclear atypia	1	Small, uniform, and regular nuclei
	2	Moderate variations in size and shape
	3	Multiple nucleoli with prominent variation
Tubule formation	1	$>75\%$ of the tumor forms tubule
	2	10–75% of the tumor forms tubule
	3	Multiple nucleoli with prominent variation

$$CancerGrade(TS^*) = \begin{cases} grade : 1, & \text{if } TS \text{ is } 3 - 5 \\ grade : 2, & \text{if } TS \text{ is } 6 - 7 \\ grade : 3, & \text{if } TS \text{ is } 8 - 9 \end{cases} \quad (1.1)$$

\* $TS \rightarrow Total\ Score$

parameter scores, cancer grade is determined as shown in Eq.(1.1). Conventionally, these three parameters are individually evaluated by manual analysis of histopathology slides, stained using Hematoxylin and Eosin (H & E), under a compound microscope. This labor-intensive procedure requires the service of an expert pathologist for each parameter’s evaluation and the final grading. The manual procedure of breast cancer grading is error-prone and has shown high interobserver disagreement (Malon *et al.*, 2012; Robbins *et al.*, 1995) with possible impact on the treatment course and outcome. Automation of this procedure has the potential to reduce the workload of pathologists, eliminate human errors, and speed up the treatment.

## 1.2.2 Molecular Subtyping of Breast Cancer

Molecular subtyping of breast cancer is based on the expression of genetic factors behind the uncontrollable proliferation of malignant cells causing tumor formation. Most

commonly accepted molecular subtypes of breast cancer are Luminal A, Luminal B, HER2-enriched, and Triple-negative/Basal-like (Al-Thoubaity, 2020). Each of these subtypes demonstrates different phenotypic expression and clinical behavior. Since the genetic factors behind these subtypes are different, the treatment required for each of them also varies from one subtype to another. St Gallen International Expert Consensus (Goldhirsch *et al.*, 2013) provides time-to-time recommendations for the targeted treatment of different molecular subtypes of breast cancer.

Determination of molecular subtype is a vital procedure for effective breast cancer treatment. A cost-effective and commonly adopted method for molecular subtyping is by immunohistochemistry (IHC) analysis (Zaha, 2014; Dabbs, 2017). In this process, the status of four key molecular biomarkers namely estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and antigen Ki67 are analyzed. This analysis is done by applying appropriate antibody reagents to the tumor tissue samples and observing the glass slides prepared this way via a microscope. Response to these antibodies indicate the presence and extent of the molecular biomarkers. The biomarker responses are assessed by pathologists to decide the molecular subtype of the tumor.

The pathology procedure for molecular subtyping is also manually done by trained pathologists. IHC slides are prepared separately for the four biomarkers by applying appropriate antibodies reagents. Typically, for each biomarker 10 hotspot regions in the slide are chosen to estimate the biomarker response. In this way, 40 hotspots per patient need to be analyzed to determine the final molecular subtype of the cancer. This labor-intensive procedure requires the service of an experienced pathologist making it costly, time consuming, and prone to inter-observer variability (Gavrielides *et al.*, 2011; Chung *et al.*, 2016).

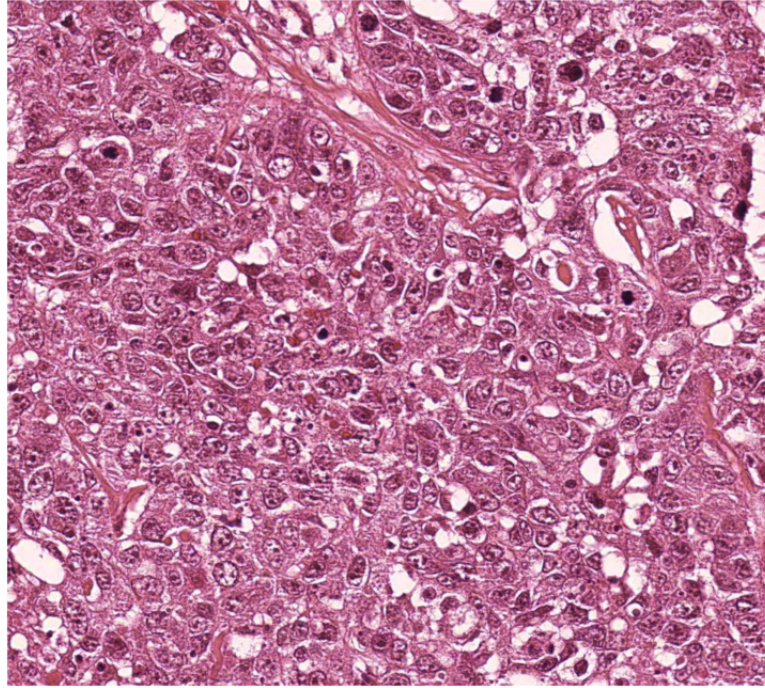


Figure 1.1: A sample high power field (HPF) slide image captured at  $40\times$  magnification of the microscope.

### 1.2.3 Automation of Pathology Procedures

Digital pathology has become an important tool in the diagnostic procedures of cancer. In this, pathology slide regions are scanned using slide scanners or camera mounted microscopes, and stored as digital images (Higgins, 2015). Such slide images captured at the maximum magnification level of the equipment are referred to as high power field (HPF) images. Figure 1.1 shows an HPF image captured at  $40\times$  magnification of the microscope used. Instead of directly analyzing the biopsy slides through a microscope, pathologists can analyze the HPF images from any location for diagnosis and prognosis. Another possibility opened up by digital pathology is the use of computational algorithms for semi-automated or fully-automated analysis of the digitized slides (Mullrane *et al.*, 2008). Pathology image analysis for different treatment aspects of cancer is an active research area for various cancer types (Thakur *et al.*, 2020; Srinidhi *et al.*, 2020).



The last 10 years have seen several research efforts to automate the grading of breast cancer using H & E-stained histopathology images created by digitizing biopsy slides. Earlier methods attempted to grade breast cancer using custom datasets with a limited number of slide images. Some of these methods followed the NGS (Dalle *et al.*, 2008) and the others did not consider the individual parameters of NGS (Doyle *et al.*, 2008; Naik *et al.*, 2008). Since small-scale proprietary datasets were used in these methods, a fair evaluation and comparison of such methods was barely possible. Automated breast cancer grading attracted attention as a relevant research problem since the launch of open contest MITOS (Roux *et al.*, 2013) targeted to address automated *mitosis detection*. A public dataset of 100 annotated H & E-stained slide images (ICPR, 2012) were made available to the research community. Since then, several methods were proposed for this challenging task using the same dataset. Automated methods for mitosis detection in the literature have been summarized by Mathew *et al.* (2020). Automation of *nuclear atypia scoring* received less research attention compared to mitosis detection. This is mainly because there were not as many public datasets or open contests available for this task. The complexity of the slide images and subjective nature of assessment criteria are other possible reasons for this trend. Results reported in the existing atypia scoring methods are also not enough to meet the requirements of clinical usage. These factors indicate the need for continued research on this task, mainly focused on tapping the potential of deep learning since deep learning algorithms like CNNs are found to be highly effective for medical image analysis.

Digital images captured from hotspot regions in IHC slides can be used to automate molecular subtyping of breast cancer. Figure 1.2 shows the digitized IHC hotspot images of the four biomarkers involved in molecular subtyping of breast cancer. These images can be analyzed using computational methods to predict the status of each biomarker for a patient and determine the molecular subtype of the cancer. Targeted treatment based on the identified subtype of the cancer increases the possibility of a better outcome. This also avoids over-treatment and reduces the financial burden and psychological trauma associated with cancer treatment.

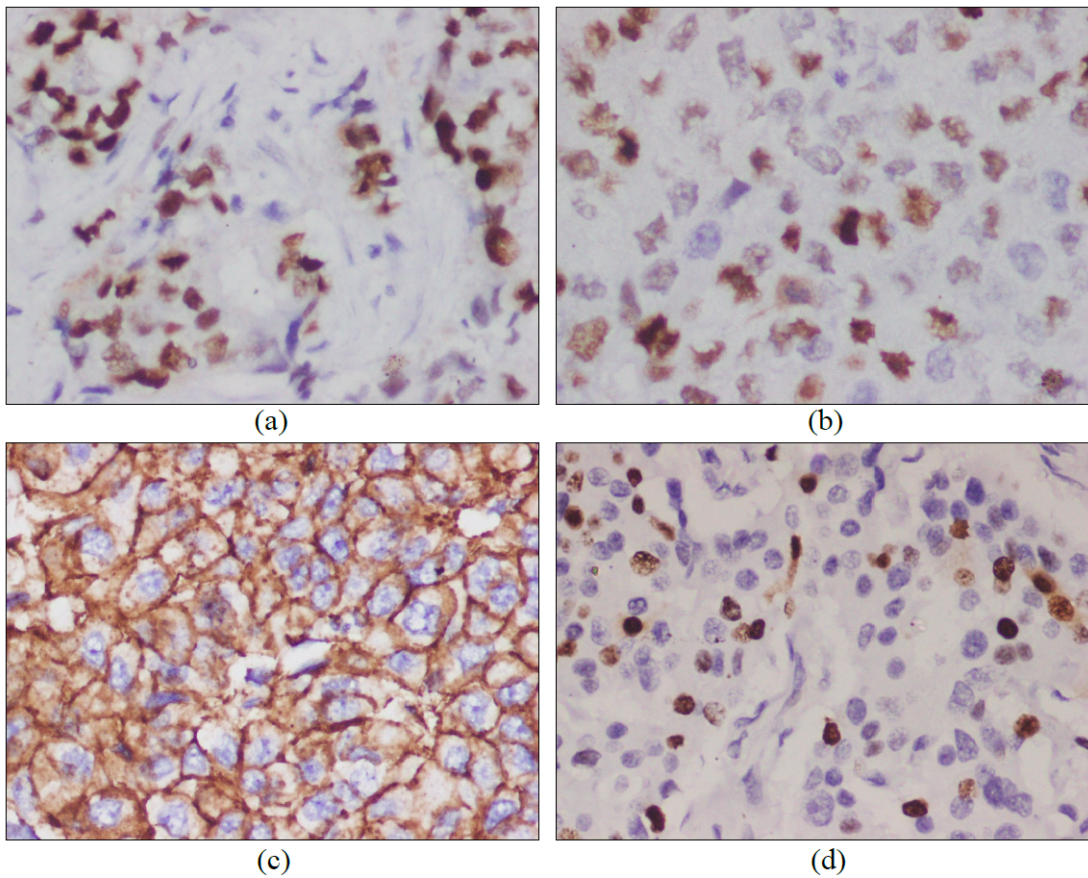


Figure 1.2: Immunohistochemistry (IHC) slide image samples of biomarkers used for molecular subtyping of breast cancer, (a) Estrogen receptor (ER) , (b) Progesterone receptor (PR), (c) Human epidermal factor receptor 2 (HER2), (d) Antigen Ki67.

### 1.3 Motivation and Problem Statement

Currently, female breast cancer is the leading cause of cancer worldwide (Sung *et al.*, 2020). Breast cancer mortality rate is also high with around 0.68 million deaths in the year 2020. Early-stage diagnosis and treatment targeting the specific molecular subtype of breast cancer has the potential to reduce the mortality rate significantly. The routine manual pathology procedures followed for cancer grading and molecular subtyping have inherent challenges related to pathologists' disagreement, time-delay of the procedure, and human-labor involved. Apart from the challenges found through the study of literature (Roux *et al.*, 2013; Malon *et al.*, 2012), the interactions with pathologists provided insights about the challenges they face in their routine clinical practice.

The manual evaluation of breast cancer grading factors namely mitotic count and atypia score are testified by them as laborious and error-prone procedures. Similarly, molecular subtyping also involves manual counting of nuclei that express certain biomarker presence. In this case, there is a requirement to evaluate a large number of hotspot regions ( $\sim 40$  hotspots per patient) in the pathology slides to determine the subtype of the cancer.

Improved availability of digitized glass slide images through public and custom datasets and advancements in artificial intelligence have inspired several research efforts on automating various pathology procedures (Niazi *et al.*, 2019). However, the study of the existing literature revealed that the current methods for automating the assessment of breast cancer grading factors, mainly mitosis count and atypia score, are not giving performance required for routine clinical application. Moreover, it is observed that the potential of deep learning has not been fully exploited in the existing methods. In the case of molecular subtyping, there are no methods found to automate this procedure using IHC slide image analysis.

### **1.3.1 Problem Statement**

Automated assessment of breast cancer grading factors has the potential to make this pathology procedure more accurate, faster, and cost-effective. Although there are several methods for automated *mitosis detection* reported in the last decade, the performance of these methods are not sufficient to meet the requirements of clinical usage. Deep learning algorithms require a large number of labeled samples to train them. Labeled datasets with sufficient sample size is a constraint for application of advanced deep learning algorithms for mitosis detection. Moreover, in the available datasets there is a large class-imbalance between the target classes of mitotic and non-mitotic figures present in the tissue images. These limitations pose barriers for the use of advanced deep neural networks for the task of mitosis detection. There is a research need to address these limitations and exploit the potential of advancements in deep learning

algorithms for automated mitosis detection.

H & E-stained histopathology slide images are structurally complex and large in dimension. Direct application of deep learning algorithms for such images is found to be less effective due to factors like computational complexity, and insufficiency of labeled training samples. The existing methods for *nuclear atypia scoring* (the second component of breast cancer grading) have failed to exploit the potential deep learning algorithms due to these factors. There are only few methods available for this clinically significant task and the performance of those feature-based learning methods show the need for continued research on this problem.

*Molecular subtyping* of breast cancer involves a significant amount of manual work by pathologists in analyzing around 40 IHC slide images per patient and evaluating various classes of nuclei, membrane etc. These factors cause time-delay and errors in the manual procedure. Even though there are methods to assess individual biomarkers involved in molecular subtyping, such methods do not lead to molecular subtyping since it requires patient-level evaluation of all four biomarkers. A consolidated method that evaluates all the four biomarkers (ER, PR, Ki67, & HER2) patient-wise is required to automate the procedure for molecular subtyping of breast cancer.

### **Research Objectives:**

The expected outcome of this thesis work is the development of novel deep learning based solutions for automated analysis of breast cancer histopathology images. Anchoring on this, the following specific objectives are identified:

1. To develop a method for automated detection of mitosis in H & E-stained histopathology images that can address the class-imbalance and sample size limitation in the datasets to enable the use of advanced deep learning algorithms for this task.
2. To develop a deep learning-based framework for automated atypia scoring of breast cancer to effectively utilize the potential of current and future deep learning algorithms.
3. To develop a deep learning-based framework for automated molecular subtyping of breast cancer using IHC image analysis of the four biomarkers' status in tumor tissues.

## 1.4 Major Contributions

The focus area of the research is application of deep learning for automated analysis of histopathology images of breast cancer. The main contributions of this thesis are summarized below:

- A novel mitosis detection method is proposed that applies an advanced convolutional neural network (CNN) architecture namely DenseNet for the first time to address this task. Applying advanced CNNs for mitosis detection was constrained by the limited sample size of the datasets and class imbalance problem in the data samples. In the proposed method, two different datasets are combined after suitable preprocessing to normalize the variations and create sufficient training samples. Class imbalance problem of the target classes is addressed by augmentation of the minority class samples in a context-preserving manner. The positive impact of combining datasets and the augmentation techniques is experimentally verified. This approach may be applied in other similar domains where multiple small datasets are available from different sources, but they are not large enough to train data hungry algorithms like CNNs independently.
- A novel deep learning-based framework is proposed for nuclear atypia scoring of breast cancer. The framework consists of three major phases namely preprocessing, deep learning, and postprocessing. The original three-class problem of slide level atypia scoring is reformulated as a six-class problem of nuclei classification for the effective use of deep learning algorithms. Subsequently, a CNN is used to classify the six classes of nuclei present in slide images. Nuclei-level analysis using the CNN approximates the manual procedure and forms a key factor in the performance. The results obtained for performance metrics precision, recall, and f1 score are improved by 13.93%, 9.89%, and 11.90% over the nearest state-of-the-art method. In addition, the problem of pathologists' disagreement and the challenges in automated nuclear atypia scoring are analyzed in detail.
- A novel classifier framework for automated molecular subtyping of breast cancer is proposed. The four protein biomarkers involved in molecular subtyping are analyzed using independent processing pipelines having preprocessing, deep learning, and post-processing stages. Consolidation of individual biomarker assessment contributes to the final determination of molecular subtype. This way the framework emulates the manual procedure of molecular subtyping through computational image analysis, at the same time reduces the human-labor and time-delay involved in the manual procedure. As a pioneering attempt for automated molecular subtyping based on IHC images, the result obtained for the proposed method is highly encouraging. The framework nature of the solution in the case of atypia scoring and molecular subtyping enables the use of different algorithms in the various stages of the framework to improve the results further.

## 1.5 Organization of this Thesis

Rest of the thesis is organized as follows:

**Chapter 2** is the literature survey carried out as part of this thesis work which encompasses the related works on the three tasks namely mitosis detection, nuclear atypia scoring, and molecular subtyping.

**Chapter 3** presents a new automated mitosis detection method developed for H & E-stained histopathology images.

**Chapter 4** presents a novel framework for nuclear atypia scoring of breast cancer using H & E images. The challenges in automated nuclear atypia scoring and the problem of pathologists' disagreement in atypia scoring are illustrated in detail.

**Chapter 5** presents a novel framework developed for molecular subtyping of breast cancer through IHC image analysis.

**Chapter 6** concludes the thesis by summarizing the findings of the thesis.

## CHAPTER 2

### LITERATURE SURVEY

In this chapter, the literature study carried out as part of the thesis work is presented. The chapter is divided into four major sections. Initially the literature on breast cancer grading and related challenges are presented. This is followed by the review of works related to the three objectives of the research work. Organization of this chapter is outlined in Figure 2.1.

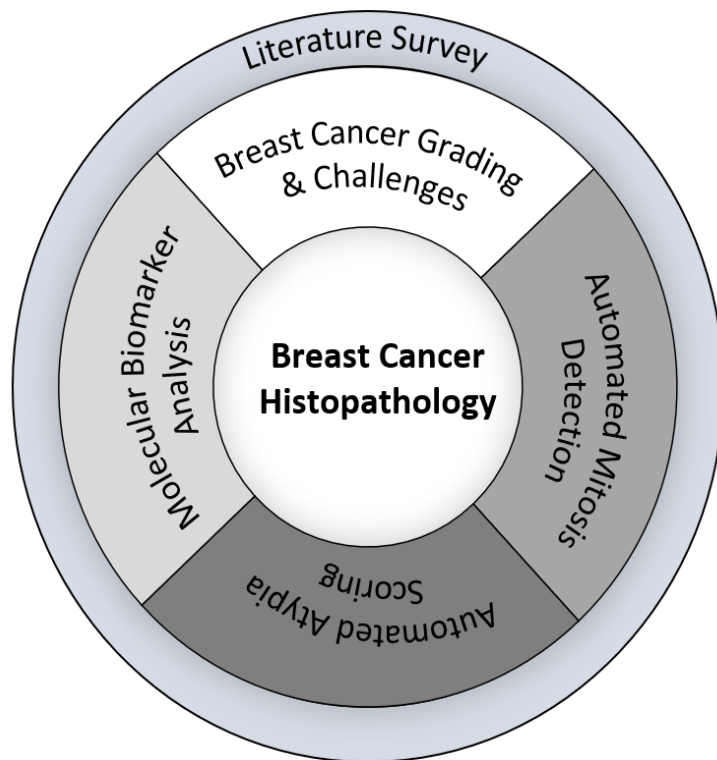


Figure 2.1: Organization of the literature survey

## 2.1 Breast Cancer Grading & Related Challenges

Studies on varying degrees of breast malignancy and their correlation with prognosis had started nearly a century ago (Greenough, 1925). Combining the outcome of all such studies and based on their own experiments, Bloom and Richardson (1957) formulated a breast cancer grading system. This was based on three factors: the tubular arrangement of cells, varying size and shape of nuclei (atypia), and frequency of mitotic figures. Elston and Ellis (2002) modified the Bloom-Richardson grading system to make the criteria more objective and well-defined to create the present NGS (Table 1.1). Pienta and Coffey (1991) studied the correlation of nuclear morphometry with breast cancer progression and concluded that even though nuclear morphometry has prognostic relevance, it cannot indicate the recurrence chances of the disease. The relation between cancer stage and the histologic grade was studied by Henson *et al.* (1991) on 22,626 cases of breast cancer to conclude that these two factors can jointly improve the prognosis of breast cancer. They also recommended the creation of a combined prognostic index using cancer stage and histologic grade.

Inter-observer variability in manual pathology procedure is a challenge in clinical practice (Nicholson *et al.*, 2004; Eaden *et al.*, 2001). This issue in the context of breast cancer grading was studied by Robbins *et al.* (1995) and they observed 80% agreement among the pathologists. Malon *et al.* (2012) studied agreement among pathologists for mitosis detection and compared pathologists' observation with an automated system. It was found that the automated method gave an encouraging performance, suggesting the viability of automated methods in such tasks. Inconsistency in nuclear atypia scoring by different pathologists is studied by Dunne and Going (2001), and they concluded that the subjective nature of atypia scoring criteria is the root cause for the scoring inconsistency. For the assessment of nuclear atypia, concordance between pathologists' independent interpretation and a reference interpretation was found to be as low as 48% in a study conducted by Elmore *et al.* (2015). With the advancement of digital pathology, increased use of automated methods is considered as a solution to reduce



pathologists' disagreement, human errors, and workload (Fuchs and Buhmann, 2011).

Efforts to automate grading of tumors based on cellular morphometric and textural parameters started towards the end of the last century (Einstein *et al.*, 1998; Wolberg *et al.*, 1995; Kaman *et al.*, 1984). In earlier days, the methods looked at tumor grading as a single task in totality rather than focusing on individual parameters of grading system. Wavelet-based multiscale image analysis is used to extract chromatin texture feature descriptors in the semi-automated method proposed by Weyn *et al.* (1998). Further, a KNN classifier is used to classify the tumors. Kronqvist *et al.* (1998) introduced optimal thresholds for various morphometric features such as the means of nuclear area, diameter, shortest axis, etc., that are used commonly in the grading of tumors. They showed that these thresholds could be used in automated grading systems based on Bloom-Richardson specifications. Cosatto *et al.* (2008) used the conventional approach of segmenting the nuclei and the resultant nuclei outlines are used to extract a set of textural and morphological features. These features are applied to classify the nuclei according to the grade using an SVM as the classifier. This method used a custom set of handpicked tumor tissue regions to train the model.

An automated grading system considering all the parameters of NGS individually is proposed by Dalle *et al.* (2008). This method is claimed to be the first such method in the literature. Individual parameter scores are computed using conventional image processing and feature extraction techniques, and finally these scores are combined to predict the overall grade of the tumor. A custom dataset of digitized slide images collected from six patients is used in this method. The results of the automated assessment showed a moderate level of matching with pathologists' evaluation. A general classification for low and high-grade breast cancer based on a large set of extracted features is proposed in the method by Doyle *et al.* (2008). This method does not follow the NGS specification for grading. A similar classification method is proposed by Naik *et al.* (2008) using graph-based extraction of nuclear features along with SVM. Dalle *et al.* (2009) used distance transform and morphological operations on the binary image

obtained by thresholding the gamma corrected R channel of tissue image to select candidate nuclei for atypia scoring. Then the nuclei are segmented using polynomial curve fitting. Size, shape, and texture features are extracted from segmented nuclei to build a Gaussian model for grade differentiation.

## 2.2 Automated Mitosis Detection

The existing methods for automated mitosis detection can be broadly categorized as the following based on the approach adopted. *i) Methods using handcrafted features, ii) Methods using deep learning, iii) Combination methods*, which combine handcrafted features and deep learning. The following subsections discuss the methods reported under these categories.

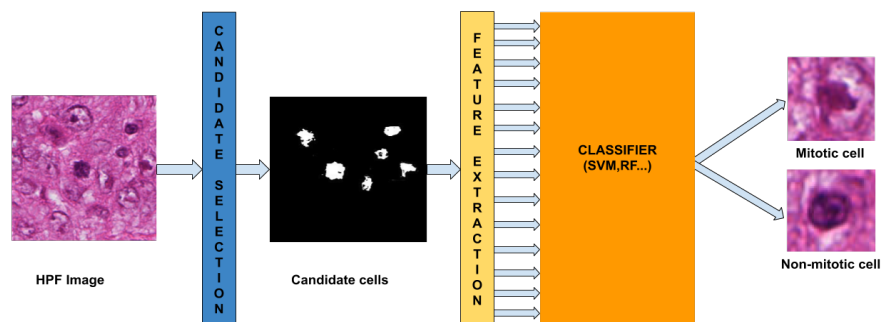


Figure 2.2: Typical workflow of methods using handcrafted features.

### 2.2.1 Methods using Handcrafted Features

Manual extraction of features from data and making machine learning algorithms to learn from these features for pattern recognition or classification is a very conventional and time-tested approach. A general structure of mitosis detection methods that adopted this pattern is presented in Figure 2.2. First, the input image is processed to identify the candidate cells/nuclei. Candidate cells consist of mitotic, non-mitotic and mimics. Subsequently, features are extracted from the candidate cells to train a classifier to discrim-

Table 2.1: Summary of the handcrafted feature based mitosis detection methods presented in this study.

Method	Approach	Dataset
Huang and Lee (2012)	Exclusive Independent Component Analysis	MITOS
Khan <i>et al.</i> (2012)	Gamma-Gaussian Mixture Model (GGMM)	MITOS
Sommer <i>et al.</i> (2012)	SVM, Random Forest (RF)	MITOS
Irshad (2013)	Laplacian of Gaussian (LoG), Active Contour Model (ACM), Decision Trees (DT)	MITOS
Tek (2013)	LoG, Cascaded AdaBoosts	MITOS
Veta <i>et al.</i> (2013)	ACM, LoG Linear Discriminant Classifier	CUSTOM
Irshad <i>et al.</i> (2014a)	Multi-spectral spatial features, SVM	MITOS
Lu and Mandal (2013)	Linear Discriminant Analysis, Bayesian Modeling, SVM	MITOS
Irshad <i>et al.</i> (2013)	SIFT, SVM, RF, DT	MITOS
Paul and Mukherjee (2015)	RF	MITOS-ATYPIA
Nateghi <i>et al.</i> (2017)	Maximum-likelihood estimation (MLE), Complete local binary pattern (CLBP), SVM	MITOS MITOS-ATYPIA
Tashk <i>et al.</i> (2013)	CLBP, SVM, MLE	MITOS
Roullier <i>et al.</i> (2011)	Multi-resolution graph-based analysis	CUSTOM
Irshad <i>et al.</i> (2014b)	Multilayer Perceptron, DT, SVM	MITOS
Nateghi <i>et al.</i> (2014)	SVM, GGMM, MLE	MITOS
Tashk <i>et al.</i> (2015)	CLBP, SVM, RF	MITOS

inate mitotic cells from the rest. While this is the general pattern observed, individual methods may deviate from this by small to large margins. For candidate cell extraction and classification, one or more of the standard algorithms or its variants are generally used.

As part of the MITOS 2012 contest, Huang and Lee (2012) proposed an algorithm named as exclusive independent component analysis (XICA) for mitosis detection. XICA finds independent bases for patterns in the training set. The similarity between the relative residuals computed from the test pattern and base patterns are measured to classify the test patterns. Inter-observer variability in manual detection of mitosis is studied in the work by Malon *et al.* (2012) and automated analysis is positioned as a potential alternative to manual detection and counting. The method proposed by Sommer *et al.* (2012) uses a random forest algorithm for nuclei segmentation and SVM classi-

fier with the Gaussian kernel to classify the candidate nuclei. For training the SVM, texture, shape and statistical features are applied. Multi-channel statistics and morphological features are used by Irshad (2013) for mitosis detection. The candidate cells are segmented using LoG, contour model, and thresholding. Further, a set of 143 features are used for classifying cells as non-mitotic and mitotic. Deviating from the general pattern, a method without object level segmentation was proposed by F. Boray Tek (Tek, 2013). Features based on color, morphology, Laplacian, and shape are used with the Adaboost classifier in this method. Veta *et al.* (2013) followed the common approach of segmentation of nuclei and feature extraction in their method for mitosis detection. Nuclei segmentation is done using the Chan-Vese level set method (Chan and Vese, 2001) and followed by a linear discriminant classifier for classification. Multispectral analysis of histopathology images is used in the methods Irshad *et al.* (2014a) and Irshad *et al.* (2014b). Compared to RGB images, multispectral images provide more chemical and anatomic features at tissue level to train the learning algorithms. Lu and Mandal (2013) proposed a three-stage method using multispectral images. The three stages are discriminative image generation by linear discriminant analysis, segmentation of candidate cells with Bayesian modeling and hybrid gray-scale morphological reconstruction, and classification of candidate cells using a multi-classifier framework. Scale-invariant feature transform (SIFT) features from H & E images are employed to train SVM and decision trees by Irshad *et al.* (2013). Relative-entropy maximized scale space is applied by Paul and Mukherjee (2015) for cell segmentation and followed by random forest for classification of mitotic and non-mitotic cells. Method by Beevi *et al.* (2016) employed localized active contour model and bio-inspired optimization to identify the candidate cells. These cells are classified using the random kitchen sink algorithm. Pixel level and object-level features are used by Tashk *et al.* (2013) in their method. Pixel level features are used to train a maximum likelihood estimation system whereas the object level features are used to train an SVM.

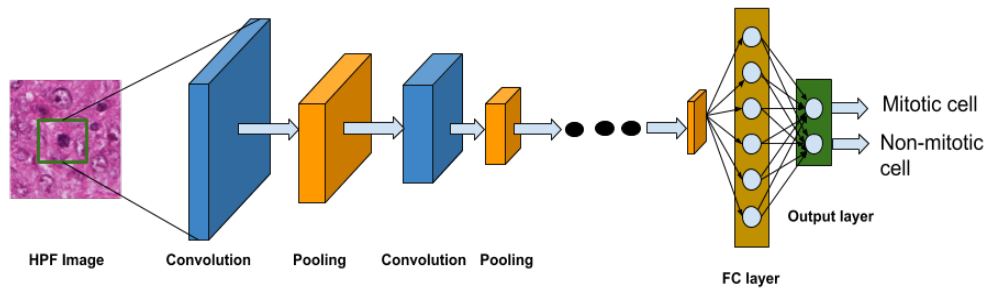


Figure 2.3: Typical workflow of methods using deep learning.

## 2.2.2 Methods using Deep Learning

The application of deep learning in medical systems is one of the top research trends nowadays. There has been widespread adoption of deep neural networks (DNN) in addressing several medical image analysis tasks (Litjens *et al.*, 2017; Anwar *et al.*, 2018). Several methods that apply deep learning for diagnostic tasks related to different types of cancers are already available in the literature (Amin *et al.*, 2020; Zhao *et al.*, 2019; Kadam *et al.*, 2019; Wang *et al.*, 2019). Not only cancer, but for many other diseases deep learning is being applied. Classification of Alzheimer's disease (Ramzan *et al.*, 2020), detection of genetic disorders (Gurovich *et al.*, 2019), etc. are just a few instances of a growing trend. In many cases, deep learning methods exceeded the performance of conventional methods. For mitosis detection, several deep learning methods started appearing in literature lately. In this section, the deep learning based methods in literature for mitosis detection are reviewed.

Convolutional neural networks (CNN) (Wu, 2017) are the most popular class of deep learning algorithms used in medical image analysis. Figure 2.3 depicts the typical workflow of a typical CNN architecture used for detection or classification problems. Input image or a selected part (sub-image) of it goes through a series of convolution and pooling layers that learn the pattern in the input image to predict the class it belongs to.

Cireřan *et al.* (2013) proposed a CNN based method for mitosis detection. In this method, each pixel is classified by considering a patch centered on that pixel. The

Table 2.2: Summary of the deep learning-based mitosis detection methods.

Method	Approach	Dataset
Cireřan <i>et al.</i> (2013)	Convolutional Neural Network (CNN)	MITOS
Albarqouni <i>et al.</i> (2016)	CNN	AMIDA
Chen <i>et al.</i> (2016 <i>b</i> )	Deep Regression Network (DRN)	MITOS
Chen <i>et al.</i> (2016 <i>a</i> )	Cascaded CNN	MITOS-ATYPIA
Wollmann and Rohr (2017)	Deep Residual Hough Voting	AMIDA
Li <i>et al.</i> (2018)	CNN	MITOS-ATYPIA
Wahab <i>et al.</i> (2017)	CNN	MITOS, TUPAC
Cai <i>et al.</i> (2019)	Regional CNN	MITOS-ATYPIA, TUPAC
Romo-Bucheli <i>et al.</i> (2017)	CNN	AMIDA
Das and Dutta (2019)	CNN, Haar Wavelets	MITOS-ATYPIA

sliding window based classification of each pixel in the HPF image is computationally intensive during training and testing. Considering the unavailability of a large dataset for deep learning, Albarqouni *et al.* (2016) proposed a framework by incorporating a crowd-sourcing layer called AggNet into CNN. Crowd-sourced image annotations are used to train the proposed CNN model. A deep regression network (DRN) with fully convolutional kernels was proposed by Chen *et al.* (2016*b*). A pre-trained model is used to offset the small number of samples in the dataset. A cascaded deep neural network with two stages was proposed by Chen *et al.* (2016*a*). The first stage is a coarse model to detect the candidate cell and the second stage differentiates the mitotic cells from its close mimics and non-mitotic cells. Computation time is significantly reduced by limiting the search space to the candidate cells. Mitosis detection from whole slide images was proposed by Romo-Bucheli *et al.* (2017). This method validates the positive correlation between mitotic activity and Oncotype DX risk score of breast cancer patients. Wollmann and Rohr (2017) used a deep residual network and Hough voting for mitosis detection. Three deep neural networks, each with different roles, are used in the method by Li *et al.* (2018). The networks include a detection network (DeepDet)

Table 2.3: Combination methods using handcrafted features & deep learning.

Method	Approach	Dataset
Malon <i>et al.</i> (2008)	Support Vector Regression (SVR), CNN	CUSTOM
Malon and Cosatto (2013)	CNN, SVM	MITOS
Wang <i>et al.</i> (2014)	CNN, RF	MITOS
Beevi <i>et al.</i> (2017)	Deep Belief Network, RF	MITOS-ATYPIA
Beevi <i>et al.</i> (2019)	CNN, RF	MITOS-ATYPIA
Saha <i>et al.</i> (2018)	CNN, ANN	MITOS MITOS-ATYPIA AMIDA

to detect the candidate cells, a verification network (DeepVer) to verify the candidates and eliminate false positives, and a deep segmentation network (DeepSeg) to segment and provide bounding boxes. Wahab *et al.* (2017) addressed the class imbalance problem by augmentation of mitotic samples and under-sampling of non-mitotic samples. A modified regional convolutional neural network (RCNN) with a ResNet backbone is used by Cai *et al.* (2019) for mitosis detection. Wavelet decomposition of the image patches ( $81 \times 81$  pixels) and using those for training a custom CNN is the approach adopted by Das and Dutta (2019) in their method. Deep learning methods for mitosis detection are summarized in Table 2.2.

### 2.2.3 Combination Methods

Some of the methods in literature combine hand-crafted features and deep learning for mitosis detection. Often it is the case that one complements the other to give an improved performance. The combination approach is first used by Malon *et al.* (2008) to detect various cellular structures like signet ring cells, mitosis, and epithelial cells. In this method, support vector regression (SVR) is used to choose the candidate elements, and the candidates are used to train the CNN. The same authors proposed another method (Malon and Cosatto, 2013) exclusively for mitosis detection. In this the CNN output is combined with color, texture, and shape features to train the SVM

classifier. Wang *et al.* (2014) used a light-weight CNN and a random forest classifier trained with hand-crafted features for mitosis detection. Use of a simple CNN results in reduced computation time for this method. Saha *et al.* (2018) proposed mitosis detection from whole slide images using a CNN and a set of 55 handcrafted features. A five-layer CNN with two fully connected layers is used in this. Transfer of weights from a pre-trained CNN model VGGNet is used by Beevi *et al.* (2019). Color variation in the image samples caused by staining differences is mitigated using color normalization. The authors reported better performance and computational efficiency over the existing methods that use raw patches. Combination methods for mitosis detection are summarized in Table 2.3.

## 2.3 Automated Nuclear Atypia Scoring

Open grand challenge MITOS-ATYPIA organized along with the International Conference on Pattern Recognition (ICPR 2014) turned out to be a landmark event in the research trajectory of automated nuclear atypia scoring. A public dataset of 600 labeled slide images, captured using two different scanners at  $20\times$  magnification, was shared with the research community to develop automated methods for atypia scoring. Most of the nuclear atypia methods reported since then used this dataset for training and evaluation. This has facilitated the performance comparison of various methods and tracking the progress of the art. One of the initial methods based on the MITOS-ATYPIA dataset is proposed by Khan *et al.* (2015). Regional covariance descriptors at the image level have been used in this method. The geodesic geometric mean of the regional covariance descriptors, computed for each non-overlapping region in the image, is defined as the global covariance descriptor for the image. Then, a geodesic kNN classifier based on Riemannian manifold of symmetric positive definite (SPD) matrices is used to assign nuclear atypia scores from the global covariance descriptors. The method proposed by Lu *et al.* (2015) first segmented the nuclei in the image us-



ing Laplacian of Gaussian (LoG) based processing of the blue-ratio image computed. Image processing techniques are then used to extract 142 textural and morphological features. An SVM is trained using these features to classify images according to the nuclear atypia score. The MITOS-ATYPIA dataset is used to develop this method as well. Maqlin *et al.* (2015) applied a restricted Boltzmann machine (RBM) with a deep neural network for nuclear atypia scoring. A contrast divergence algorithm is used to train RBM in each layer of the model independently. The stacked RBMs thus form a deep belief network (DBN). The DBN is fine-tuned with the use of a backpropagation algorithm. The method used a subset of only 80 slide images from the MITOS-ATYPIA dataset. Multi-scale descriptors computed from segmented nuclei are the basis of the method proposed by Moncayo *et al.* (2015). These descriptors are clustered by the k-means algorithm and used as atoms of a learned dictionary. Histogram based features of these descriptors are utilized to train an SVM or a bank of binary classifiers to grade each atom in the dictionary using the score labels associated. This method is developed and evaluated using breast cancer images from 'The Cancer Genome Atlas' (TCGA) database. Wan *et al.* (2017) adopted the approach of nuclei segmentation followed by feature extraction at pixel, object, and semantic levels to train multiple SVMs that classify nuclear atypia according to the grade. A hybrid active contour method consisting of boundary and region information is used for nuclei segmentation. Semantic features are extracted using a CNN. A custom dataset of H & E images is used for developing this method.

Several learning algorithms are available for breast cancer diagnosis and classification (Tariq *et al.*, 2020; Houssein *et al.*, 2020). Recently, deep learning algorithms like CNNs have attracted much attention (Ting *et al.*, 2019) due the superior performance observed for many tasks. However, for nuclear atypia scoring CNNs are scarcely utilized. A hybrid CNN model with multiple image resolutions is used by Xu *et al.* (2017) for atypia scoring. This model consists of three single resolution CNNs that work on different image resolutions  $10\times$ ,  $20\times$ , and  $40\times$  to independently score nuclear atypia. Finally, the individual scores are combined using plurality voting. Fisher discriminant

Table 2.4: Summary of the major works in literature for automated atypia scoring

Method	Approach	Dataset
Das <i>et al.</i> (2018)	Region covariance descriptors, Dictionary learning	MITOS-ATYPIA
Das <i>et al.</i> (2020a)	Kernel-based fisher analysis, Batch mode active learning	MITOS-ATYPIA
Das <i>et al.</i> (2019)	Riemannian manifold, Fisher discriminant	MITOS-ATYPIA
Lu <i>et al.</i> (2015)	Laplacian of Gaussian, Texture features, SVM	MITOS-ATYPIA
Khan <i>et al.</i> (2015)	Regional covariance descriptors, Geodesic kNN classifier	MITOS-ATYPIA
Rezaeilouyeh <i>et al.</i> (2016)	Deep learning, Shearlet coefficients	MITOS-ATYPIA
Xu <i>et al.</i> (2017)	Deep learning, Multi-resolution CNNs	MITOS-ATYPIA

analysis on Riemannian manifold is used in the method by Das *et al.* (2019), for nuclear atypia scoring. This method also uses the geodesic mean of region covariance descriptors for the kernel-based fisher analysis. Recently, a variant of this method with batch mode active learning (Das *et al.*, 2020a) is found to give superior performance over the kernel-based Fisher discriminant analysis. The same authors proposed another method (Das *et al.*, 2018) based on sparse coding and dictionary learning. This method also used the Riemannian manifold on region covariance descriptors. The dictionary learning task is mapped to a highly discriminative high-dimensional Hilbert space resulting in the superior performance of the method. This method has reported the best result for nuclear atypia scoring so far. In the method proposed by Gandomkar *et al.* (2019), nuclear atypia scoring is based on the cytological criteria estimated by pathologists (nuclei size, nucleoli size, etc.) as well as the features such as first-order statistics features, Haralick features, etc., extracted using image processing. Scores assigned by two independent regression models trained on cytological features and extracted features are combined by a third regression model to predict the final atypia score. In the CNN-based method, Rezaeilouyeh *et al.* (2016) combined handcrafted features such as phase and magnitude of shearlet coefficients with the original image and used as the in-

put to a CNN. The additional features provided are found to give an improved accuracy compared to the sole image input to CNN. Table 2.4 summarizes the major works on automated nuclear atypia scoring from the literature.

## **2.4 Automated Analysis of Molecular Biomarkers**

Automated molecular subtyping of breast cancer is a patient-level procedure that requires the samples of all the biomarkers (ER, PR, Ki67, and HER2) from the same patient. Currently there are no methods in the literature that perform a collective assessment from all these biomarkers to predict the cancer subtype for a patient. However, there are several methods that assess one or two of these biomarkers using image analysis techniques. Such methods are summarized in this section of the literature study.

### **2.4.1 ER and PR**

According to a recent cohort study (Al-Thoubaity, 2020), the most common subtype of breast cancer is Luminal A (58.5%). This subtype is characterized by positive status of hormone receptors ER and PR. Many researchers have attempted to automate the analysis of these hormone receptors. Responses of these biomarkers to IHC reagents are more-or-less identical and hence the method for automated analysis can be similar for both. As a result, many methods in the literature have addressed the automated assessment of ER and PR together. An early study on the feasibility of image analysis for hormone receptor status prediction was carried out by Mofidi *et al.* (2003) with the help of edge-based features in a semi-automated way. The obtained results highly correlated with the manual assessment of hormone receptors as well as the objective measurements like percentage of positive nuclei. *ImmunoRatio* (Tuominen *et al.*, 2010) is a publicly available application for performing quantitative analysis of ER and PR. Vijayashree *et al.* (2015) compared manual and automated quantification of hormone receptors ER

and PR in the case of breast carcinoma. Manually evaluated HScore (McCarty *et al.*, 1986) and Allred Score (Allred *et al.*, 1998) are found to correlate with the result of auto-evaluation by *ImmunoRatio*. Method by Oscanoa *et al.* (2016), segmented the nuclei in ER images using the features such as shape and size. This is followed by the fuzzy C-means algorithm to classify them into ER +ve and ER -ve. This method used the publicly available Stanford University TMA dataset (SU, 2001). The method gave sensitivity 95.7% and specificity 93.2%. In one of the earlier methods proposed by Rexhepaj *et al.* (2008), a fully automated algorithm is applied to quantify both ER and PR biomarker responses. The images that are scanned from tissue microarrays of breast cancer are used as the dataset. Optimal thresholds of ER and PR are determined using the random forest classifier for survival analysis. As a deviation from commonly used IHC image analysis, Chaudhury *et al.* (2014) applied the features extracted from breast tissue MRI images to predict ER status. They used textural kinetic features from various tumor subregions in MRI images for ER classification. Chang *et al.* (2016) also used dynamic contrast-enhanced MRI images for determination of ER and HER2 status. In the method proposed by Mouelhi *et al.* (2014), ER status is evaluated using color deconvolution and morphological operation. The nuclei present in the IHC images of tumor regions are segmented and separated to estimate the number of positive and negative nuclei. A CNN based cell classification method for ER from whole slides images was proposed by Jamaluddin *et al.* (2018). This method detected the tumor cells in the whole slide images of breast cancer patients and classified them into four classes such as weak, moderate, and strong cells with respect to ER response on IHC staining. Abubakar *et al.* (2019) analyzed that quantitative measures of the molecular biomarkers combined with other factors such as cancer grade, lymph node involvement, tumor size, and age provided more prognostic information than categorical status of these biomarkers in Luminal breast cancers. A deep learning framework for ER and PR scoring was proposed by Saha *et al.* (2020). The framework contains a segmentation component followed by a scoring component. The segmentation component takes IHC biomarker images of ER/PR as the input and segments the nuclei using a deep CNN.

These nuclei are then passed to the scoring component that classifies the nuclei as immunopositive or immunonegative. H-Score (McCarty *et al.*, 1986) of a slide image is computed using the counts of weak, moderate, and intermediate nuclei.

#### **2.4.2 HER2**

The biomarker HER2 accelerates the growth and division of cells in tumor sites leading to an uncontrollable growth of tumors. While the hormonal receptors ER and PR are present in the tumor cell nuclei, HER2 is found in the cell membrane (Perez *et al.*, 2014). As a result, the IHC response of HER2 is visible as brownish circular layers attached to the cell membrane around the nuclei. IHC analysis results in three possible states: ‘HER2 negative’, ‘HER2 positive’, and ‘HER2 equivocal’. The equivocal status indicates that HER2 status is not clearly identified by IHC analysis and further a fluorescent in situ hybridization (FISH) analysis is required for such cases. A comparative study of HER2 score computation using FISH and IHC was conducted by Yaziji *et al.* (2004). The study suggested IHC as an efficient approach for evaluating HER2. Lloyd *et al.* (2010) studied the reliability of image analysis algorithms for the assessment of ER and HER2. They used two commercially available algorithms for this analysis and the results were compared with manual scoring by pathologists. It is found that the results of the algorithmic analysis matched with the manual scoring by pathologists. They observed that the quality assurance of the region selection process for image analysis has an influence on the accuracy of the results. Final suggestion is to use algorithmic analysis as a supplement to manual evaluation. Comparative study between IHC image analysis and FISH scoring by Ayad *et al.* (2015) suggested IHC image analysis as a potential alternative to costly and time-consuming FISH test provided the performance of equivocal cases is refined further. A multistate method for HER2 scoring from FISH images was proposed by Raimondo *et al.* (2005). The method used various techniques like top hat transform, distance transform, and marked watershed transform in the method pipeline. Considering the cost effectiveness of IHC testing and data availability, image

analysis based IHC acquired more prominence in the research community. A web application named *ImmunoMembrane* was developed by Tuominen *et al.* (2012) for the assessment of HER2 from IHC images. Based on the intensity and completeness of cell membranes, a quantitative score is generated for HER2 images under analysis. The samples are classified into one of the scoring classes as per American society of clinical oncology (ASCO) guidelines (Wolff *et al.*, 2007). Method by Hall *et al.* (2008) used a membrane isolation algorithm followed by quantitative analysis of the separated membrane to assess HER2 score from the IHC image. Results of the automated method are found to be similar to that of manual assessment and FISH test. The authors suggest that image analysis based HER2 scoring as an alternative to manual assessment and FISH test especially for the cases in equivocal range. Skaland *et al.* (2008) used basic image processing techniques like color deconvolution, thresholding, segmentation etc., to segment membrane bound IHC stain for HER2 scoring. The segmented membrane regions are quantitatively analyzed to assign scores for HER2 response. The result of the automated scoring is correlated with modified FISH scores. Consequently, the authors suggested IHC based automated scoring as a cost-effective supplementary tool for HER2 scoring. A deep learning-based method for HER2 scoring was proposed by Vandenberghe *et al.* (2017). The nuclei in the HER2 response images are detected using color deconvolution and watershed algorithm. Image patches of size  $44 \times 44$  are extracted based on the nuclei to train and test a deep learning algorithm. The CNN used in this method consists of three convolution layers and one fully connected layer for the classification of nuclear patches. The results obtained by this method showed 83% matching with the assessment of a pathologist. In another deep learning method proposed by Pitkäaho *et al.* (2016), the image patches of size  $128 \times 128$  extracted from HER2 images are used to train a CNN. An HER2 slide image is assigned with a score based on the classification pattern of patches extracted from it by the CNN classifier used. Saha and Chakraborty (2018) proposed a deep learning framework named Her2Net for segmenting the cell membranes from IHC images of breast cancer and HER2 scoring based on the segmented membranes. The CNN consists of convolution

and deconvolution segments along with the trapezoidal long short-term memory (TLSTM) units to improve the segmentation performance.

### 2.4.3 Ki67

Antigen Ki67 is found in cells under the various stages of division (Gerdes *et al.*, 1991). The number of nuclei with Ki67 presence is an indicator of tumor growth rate. Hence, Ki67 is also one of the biomarkers that has an impact on the prognosis and treatment plan of breast cancer. Moreover, Ki67 proliferation index is an essential factor in molecular subtyping of breast cancer. Automated assessment of Ki67 status from IHC images has been attempted by many researchers in the past. Abubakar *et al.* (2016) developed an automated protocol for Ki67 scoring based on the features provided in the Ariol system for microscopy image analysis. The protocol involved the detection of nuclei present in IHC images and training classifiers using these nuclei. The automated protocol showed a good correlation with computer assisted visual scoring done on the same set of tissue microarray. Automated quantification of Ki67 from the IHC images of nasopharyngeal carcinoma was proposed by Shi *et al.* (2016). The method pipeline involves the preprocessing of images, feature extraction, clustering based segmentation of immunopositive nuclei, separation of touching nuclei, and quantification. The result of automated quantification matched with the manual process by pathologists. A comparison of Ki67 labeling index by visual assessment and digital image analysis is carried out by Zhong *et al.* (2016). This study showed a perfect correlation with the results obtained by both approaches on a cohort study of 155 breast cancer cases. Comparison of Ki67 hotspot selection from whole slide images (WSI) of meningiomas using manual, semi-automated, and automated approaches is done by Swiderska *et al.* (2015). The results of all the three approaches have shown a good level of agreement.

An integrated dictionary learning based framework for automated Ki67 counting in neuroendocrine tumor images is proposed by Xing *et al.* (2013). The framework consists of three stages. The detection and segmentation of cells in Ki67 images are

Table 2.5: Summary of the works related to IHC image analysis to assess the different cancer biomarkers

<b>Biomarker</b>	<b>Method</b>	<b>Approach</b>	<b>Dataset</b>
ER	Rexhepaj <i>et al.</i> (2008)	Random forest classifier	Custom TMA dataset
	Tuominen <i>et al.</i> (2010)	Color deconvolution, Adaptive thresholding	Custom IHC dataset
	Mouelhi <i>et al.</i> (2014)	Color deconvolution, Morphological operations	Custom IHC dataset
	Oscanoa <i>et al.</i> (2016)	Feature based, Fuzzy C-means	Public TMA dataset
	Jamaluddin <i>et al.</i> (2018)	Adaptive thresholding, Deep learning	Custom WSI images
	Saha <i>et al.</i> (2020)	Deep learning	Custom IHC dataset
PR	Rexhepaj <i>et al.</i> (2008)	Feature based, Random forest	Custom TMA dataset
	Tuominen <i>et al.</i> (2010)	Color deconvolution, Adaptive thresholding	Custom IHC dataset
	Saha <i>et al.</i> (2020)	Deep learning	Custom IHC dataset
HER2	Skaland <i>et al.</i> (2008)	Color deconvolution, Thresholding	Custom IHC dataset
	Tuominen <i>et al.</i> (2012)	Colour Deconvolution based segmentation	Custom TMA dataset
	Pitkäaho <i>et al.</i> (2016)	Deep learning	Warwick dataset
	Vandenberghe <i>et al.</i> (2017)	Deep learning	AstraZeneca BioBank, Custom images
	Saha and Chakraborty (2018)	Deep learning	Warwick dataset
Ki67	Tuominen <i>et al.</i> (2010)	Color deconvolution, Adaptive thresholding	Custom IHC dataset
	Konsti <i>et al.</i> (2011)	Color deconvolution, Segmentation	Custom TMA dataset
	Xing <i>et al.</i> (2013)	Dictionary learning	Custom IHC dataset
	Niazi <i>et al.</i> (2014)	Graph cuts, Difference of Gaussians	Custom IHC dataset
	Abubakar <i>et al.</i> (2016)	TMA specific classifier algorithms	Public TMAs dataset
	Shi <i>et al.</i> (2016)	Clustering of local correlation features	Custom IHC dataset
	Saha <i>et al.</i> (2017)	Deep learning	Custom IHC dataset
	Lakshmi <i>et al.</i> (2019)	Deep learning	Custom IHC dataset
Note: No methods are currently found in literature for automated molecular subtyping of breast cancer through IHC image analysis.			

performed in stage 1. A dictionary-based learning is used to segregate tumor and non-tumor cells in stage 2. In stage 3, the tumor cells are further classified into immunopositive or immunonegative for Ki67 indexing using a color histogram-based classifier. A method to assess Ki67 expression and study its prognostic value for breast cancer is



proposed by Konsti *et al.* (2011). IHC stained tissue microarray images (TMA) of 1931 patients are used to conduct the study. The color deconvolved TMA images are thresholded to obtain hematoxylin and diaminobenzidine masks. These masks are merged by pseudocoloring and the extent of Ki67 response is measured. An algorithm based on perceptual clustering for hotspot detection in Ki67 response images is proposed by Niaz *et al.* (2014). IHC images of neuroendocrine cancer are used in this method. Graph cuts and difference of Gaussian are applied to detect the cells from Ki67 images. Pathologists' way of hotspot detection is mimicked by particle swarm optimization along with message passing clustering. For the first time, Saha *et al.* (2017) exploited the potential of deep learning for hotspot detection and proliferation scoring of Ki67 in breast cancer images. The method uses a gamma mixture model with the expectation maximization for seed point detection. This is followed by seed-based patch extraction for deep learning. The patches are classified into Ki67 immunopositive or immunonegative by the CNN used. The results obtained for precision, recall, F-score score are 0.93, 0.88, and 0.91 respectively for patch level classification. Slide level or patient level Ki67 proliferation status is not reported in the method. Lakshmi *et al.* (2019) used U-Net based deep learning architecture for the segmentation of immunopositive and negative tumor nuclei from Ki67 images of bladder cancer. Connected component analysis is applied to estimate Ki67 proliferation index from segmentation output. Table 2.5 summarizes the recent methods in the literature for assessment of individual biomarkers related to breast cancer.

## 2.5 Summary

The review of the existing literature led to many valuable observations and identification of research gaps in automation of mitosis detection, nuclear atypia scoring, and molecular subtyping of breast cancer. They are summarized below.

Research on automated mitosis detection has been active for over a decade. How-

ever, the performances of the existing methods are still far from the requirements for clinical usage. The diversity in size, shape, and textural characteristics of mitotic cells along with their close similarity with apoptotic cells make this problem a challenging one. The potential of deep learning in medical image analysis has not been exploited fully for this task. The major reasons for this is the lack of sufficiently large datasets required for advanced deep learning algorithms and the class imbalance between the nuclei of the target classes that affects the learning by deep learning algorithms. Research efforts are needed to address these issues to improve the performance of automation.

Despite several methods reported, the performance of automated atypia scoring achieved so far is not yet sufficient to apply in clinical practice. Many methods reported high performance on small custom datasets that make the reliability of such methods questionable and comparison with methods that use public datasets difficult. Among all the methods reported using the publicly available MITOS-ATYPIA dataset, the method by Das *et al.* (2018) has shown the best performance so far with precision 0.7694, recall 0.7971, and F1 score 0.7815 for combined image sets from both scanners. These results are not sufficient for applying automated atypia scoring in clinical practice, and also point to the need for further research. Another important observation from the literature review is about the use of deep learning for atypia scoring. The potential of deep learning algorithms like CNNs are not sufficiently explored for automated nuclear atypia scoring in spite of the revolutionary changes brought by such algorithms in several medical image analysis tasks (Litjens *et al.*, 2017; Pang *et al.*, 2020).

In the extensive survey carried out on molecular biomarker assessment, it is observed that the existing methods focus on quantification and status prediction of one or at the most two biomarkers of breast cancer that are required for molecular subtyping. Some of these methods are developed using proprietary or public tissues microarray (TMA) datasets. Other methods used custom IHC slide image datasets obtained from pathology labs. None of these datasets contains images of all the four biomarkers required molecular subtyping at a patient-level. Consequently, there are no methods

found in the literature to assess all the four biomarkers to perform the higher-level task of determining the molecular subtype of breast cancer. The novel deep learning-based framework proposed in this thesis work aims to address this research gap.

The following chapters 3, 4, and 5 elaborate the proposed methods to address the major research gaps identified for *mitosis detection*, *nuclear atypia scoring*, and *molecular subtyping* of breast cancer respectively.



## CHAPTER 3

# AUTOMATED MITOSIS DETECTION IN HISTOPATHOLOGY IMAGES

There are three parameters in breast cancer grading namely mitosis count, nuclear atypia, and tubule formation. Among them, mitotic cell counting is the most challenging task for pathologists where they have to identify and count the mitotic cells which typically will range from 0 – 5 in a high-power field view among the hundreds of other non-mitotic figures. The two major challenges in developing effective deep learning-based methods for mitosis detection are lack of sufficiently large datasets, and class imbalance between mitotic and non-mitotic cells in slide images. In this chapter, a new approach and a method based on that are proposed to address these challenges. High training data requirement of the advanced deep neural network is met by combining two datasets from different sources after a color-normalization process. Class imbalance is addressed by the augmentation of the mitotic samples in a context preserving manner. Finally, an advanced classifier CNN is used to classify the candidate cells into the target classes. We have used the publicly available datasets MITOS-ATYPIA and MITOS for the experiments. The proposed method outperforms most of the recent deep learning-based methods that are based on independent datasets and at the same time offers adaptability to combination of datasets from different sources.

### 3.1 Introduction

In the process of cancer treatment, grading plays a crucial role. Grade of malignant tumor indicates how much they resemble the parent tissue, i.e. degree of differentiation. Well differentiated tumors have a better prognosis as they are less aggressive than the poorly differentiated tumors. Aggressiveness indicates how fast the tumor is growing and how likely it can spread to other parts of the body. Mitotic count is a predominant objective parameter in breast cancer grading as per Nottingham Grading System (NGS) (Table 1.1). In NGS, a pathologist observes the H & E-stained histology slides through

a microscope and manually assigns scores to each of the parameters. Such a manual grading procedure is laborious and error-prone due to a large number of cells per high power field (HPF) and varying appearance of the cells under mitosis (Paul and Mukherjee, 2015). These aspects lead to high inter-observer variability (Fuchs and Buhmann, 2011) in pathological findings. Moreover, in developing and under-developed countries, pathology services are scarcely available in rural areas. These countries face an acute shortage of experienced pathologists, which is a hindrance to early diagnosis of cancer. Consequently, cancer death rates are high in such countries (Bray *et al.*, 2018). In such a context, automated detection methods can help in faster diagnosis and accurate grading to decide the appropriate treatment plan, even from distant places. This can effectively bring down cancer death rates. Automatic mitosis detection is one such step towards developing a completely automated cancer grading system for breast cancer.

### **3.1.1 Challenges in Automated Mitosis Detection**

Automating mitosis detection through histopathology image analysis has the potential to overcome these challenges associated with manual process. This can also make the procedure faster and easily accessible. Automated mitotic detection has attracted a lot of research interest in the recent past, driven by some of the open challenges organized by scientific agencies (Mathew *et al.*, 2020). Results of the reported methods have shown gradual improvement over the years. However, the performances of these methods are often specific to the datasets used and they may not perform the same way with a new dataset. This is primarily due to the variations across the datasets resulting from staining differences and acquisition setting used. Automated mitosis detection has got a set of challenges to address. Mitotic cells vary in their size, shape, and texture based on the phase of the cell division such as prophase, metaphase, anaphase etc. Such variations make it difficult for the algorithms and even human observers to distinguish them from dead cells (apoptotic cells) and other cellular structures that mimic the appearance of mitotic cells. Figure 3.1 demonstrates the variations among the mitotic cells and their

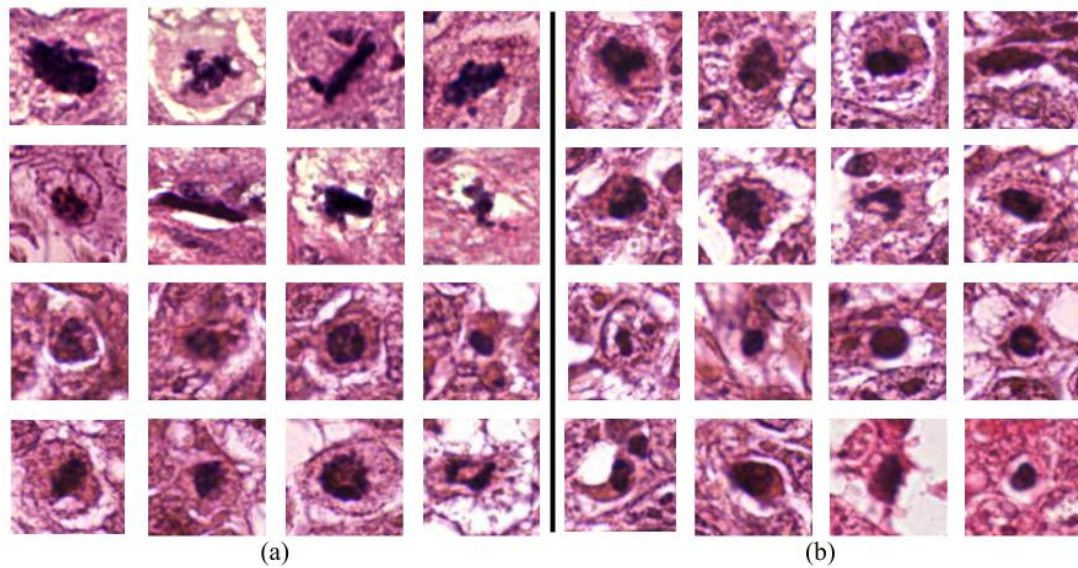


Figure 3.1: Challenges in mitosis detection related to appearance of cells. (a) Shape and size variations among mitotic cells, (b) Non-mitotic cells or structures that resemble mitotic cells in appearance.

visual similarity with other cellular structures. Another problem is that typically in HPF images mitotic cells are far less in number compared to non-mitotic cells. This causes class imbalance problem in learning based approaches. In addition to that, staining variations, make and configuration of the acquisition devices etc., introduce differences in nature of the images from different datasets that adversely impact the performance of the detection methods. As a result, a method that works well on one dataset may not perform the same way on another dataset. Methods that are resistant to such dataset variations are needed for usage in clinical practice. Class imbalance between the target objects and lack of sufficiently large training dataset are major challenges for deep learning-based mitosis detection methods.

In this chapter, a new deep learning-based method for mitotic cell detection from H & E-stained histopathology images of breast cancer is proposed. An advanced deep convolutional neural network (CNN) is used as a major component in the method pipeline. In this patch-based approach, the HPF images are divided into small-sized non-overlapping image patches to train the model. Data augmentation is applied on the

mitotic patches to remove the acute class imbalance between mitotic and non-mitotic samples used in training. A data augmentation technique referred to as context preserving data augmentation (CPDA) is applied for patch-based training of the model. The public dataset MITOS-ATYPIA (ICPR, 2014) is used as the base dataset in the proposed method. To meet the high data requirement of the deep learning algorithm used, the MITOS-ATYPIA dataset is combined with the MITOS dataset. The color variations in the image samples from these two datasets are reduced by a color-normalization process. The model trained with the combined dataset gave improved performance over the model trained with the base dataset alone.

Rest of this chapter is organized as follows. In Section 3.2, the method proposed for mitosis detection is explained in detail. Section 3.3 elaborates the experiments conducted, presents the observed results, and compares the results with the state-of-the-art methods in literature. This section is concluded with a discussion on the highlights of the proposed method in the context of observed results.

## 3.2 Methodology

In this section, the proposed method for mitosis detection is explained in detail. Figure 3.2 gives an overview of the proposed method. The upper block is the training pipeline which consists of preprocessing stages and CNN training, whereas the lower block shows the testing pipeline using the trained model. The four major stages in the proposed method are, *i) Image color normalization, ii) Candidate cell detection, iii) Context preserving data augmentation, iv) CNN training & evaluation*. An input HPF image passes through all these stages in the process of detecting mitotic cells. First two stages are common for training and testing pipelines. In the third stage, image patches are extracted in both pipelines, whereas augmentation is performed only for training. The individual stages are elaborated in the following subsections.



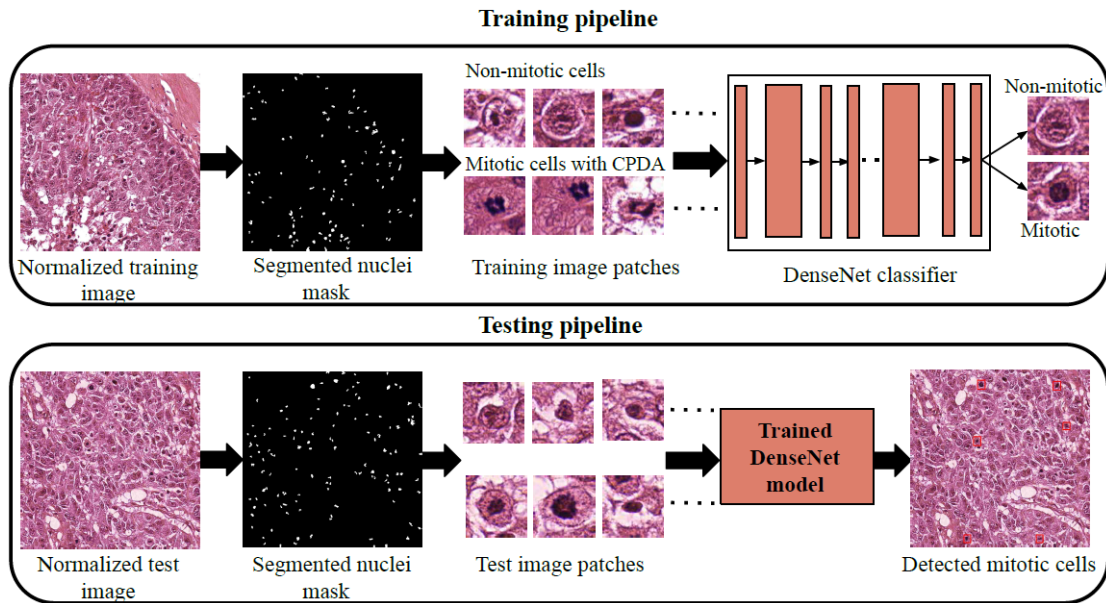


Figure 3.2: A graphical outline of the proposed method that involves a training pipeline and a testing pipeline. In the training pipeline, context preserving data augmentation (CPDA) is applied for the mitotic cell patches. The output of the training pipeline is a trained CNN model employed in the testing pipeline to classify the cell images.

### 3.2.1 Image Color Normalization

One of the major challenges in pathology image analysis is the color variations in the images resulting from factors like non-uniform staining, scanner make and configuration, illumination etc. Hence, as a pre-processing step, histopathology images are color-normalized to mitigate these variations and to transform these images to a common color level. There are many color normalization techniques in literature (Khan *et al.*, 2014; Li and Plataniotis, 2015; Vahadane *et al.*, 2016). A normalization technique known as Reinhard normalization (Reinhard *et al.*, 2001) is applied to normalize the HPF images from different scanners and datasets. This method converts the color characteristics of an HPF image to that of a desired reference H & E image used. Figure 3.3 shows the color normalization of HPF images from two different scanners Aperio (images a & b) and Hamamatsu (images c & d). The color characteristics of the images procured using these two scanners are visibly different. Color normalization of

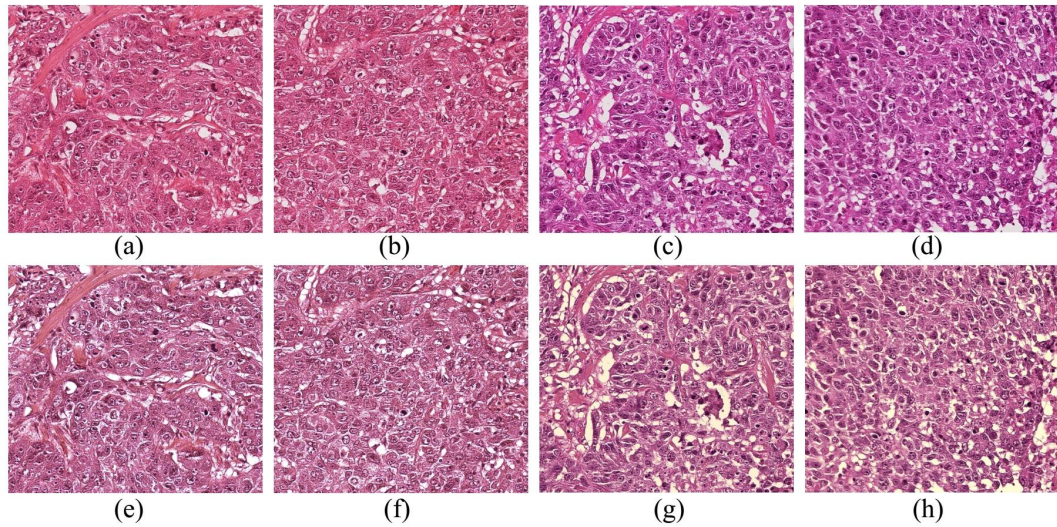


Figure 3.3: Sample outputs of the Reinhard color normalization of H & E images. Images (a, b, c, & d) are original images from two different scanners of the MITOS-ATYPIA dataset (ICPR, 2014) and (e, f, g, & h) are the corresponding color-normalized images.

these images result in uniformity of color as shown in Figure 3.3 (e, f, g, & h). The output images have a uniform color pattern compared to the input images. The illustrated image samples are part of the same dataset but captured using different scanners. Images from two different datasets are found to have more color variations and hence the normalization process is an effective step in such cases also.

### 3.2.2 Candidate Cell Detection & Segmentation

The next step in mitotic cell detection is to detect all the candidate cells in the HPF images. The set of candidate cells are identified by nuclei segmentation of the normalized HPF images from the previous step. Candidate cell detection helps to avoid the processing of unwanted regions like necrosis, fat globules, and empty regions in the slide images where target cells are not present. Nuclei are the most prominently visible component of a cell. So, the candidate cell detection is based on the detection of the nuclei.

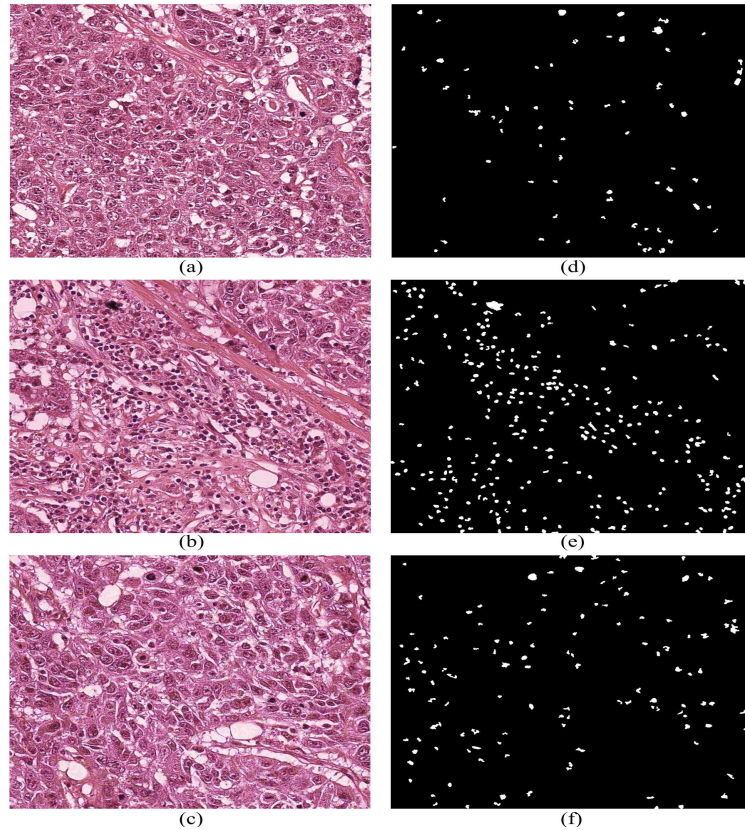


Figure 3.4: Candidate cell segmentation of HPF images. Images a, b, c are the normalized HPF images and d, e, f are the corresponding segmentation output.

We have used a nuclei detection and segmentation method specifically for H & E histopathology images, proposed by Al-Kofahi *et al.* (2009). This method consists of three major stages namely i) Image binarization, ii) Nuclei seed detection and Initial segmentation, and iii) Segmentation refinement with  $\alpha$ -Expansions and Graph Coloring. H & E staining of tumor tissues results in differential color binding where the nuclei are seen in dark purple color and cytoplasm and other cell structures in pink color. Through a color-deconvolution process on the H & E image, a nuclear channel image  $I_N(x, y)$  is extracted.  $I_N(x, y)$  is a grayscale image with nuclei appearing with different intensity compared to the remaining pixels.  $I_N(x, y)$  is further processed through the three stages to segment the nuclei. First,  $I_N(x, y)$  is binarized using the minimum error thresholding algorithm (Fan, 1998) and the fast max-flow/min-cut algorithm (Boykov and Kolmogorov, 2004a). In the second stage, a multi-scale Laplacian-of-Gaussian

based approach is used to detect the nuclei seeds and obtain an initial segmentation. This is further refined using  $\alpha$ -Expansions (Boykov and Kolmogorov, 2004a) and a graph coloring algorithm in the final stage to give a more accurate segmentation of the nuclei present in H & E images.

Once the nuclei present in the image are segmented, patches are extracted based on the presence of cells and labeled as mitotic or non-mitotic using the ground truth annotations in the dataset. HPF image regions without the presence of any cells are excluded in this process. Figure 3.4 shows the output of the candidate cell segmentation on three sample HPF images. The number of candidate cells in HPF images varied from a few dozens to a few hundreds. Out of these, only a few cells (1–5) normally belong to the mitotic class. This is the reason behind severe class imbalance between the two target classes. Our proposed method relies on detection rather than accurate segmentation of nuclei due the patch-extraction used in the following step. Table 3.1 shows the results of the nuclei detection algorithm (Al-Kofahi *et al.*, 2009) on the datasets used in our method. The detection rate is computed only for mitotic cells since the datasets contain ground-truths for this class alone. The average detection rate is 98.60% which indicates that the algorithm is effective in detecting most of the mitotic nuclei present in the HPF images.

Table 3.1: Nuclei detection rate (mitotic) given by the detection algorithm (Al-Kofahi *et al.*, 2009) used in the proposed method.

<b>Dataset fold</b>	<b>Detection rate</b>
Fold 1	97.28
Fold 2	98.89
Fold 3	98.45
Fold 4	100.0
Fold 5	98.38
<b>Average</b>	<b>98.60</b>

### 3.2.3 Context Preserving Data Augmentation

Deep learning algorithms require large amount of data samples for training since the algorithm learns the discriminating features of the target class objects through these samples. Often, the dataset may not be as large as required or there can be class imbalance between the target classes. In the proposed method, a patch-based approach is used where patches of a desired size ( $96 \times 96$  pixels in this case) are cropped from the HPF images to use for training. In this way, a large number of image patches are extracted from the images in the dataset. However, the class-imbalance between mitotic and non-mitotic cells in the HPF images causes non-mitotic patches to out-number the mitotic patches by a huge margin. This can negatively impact the feature learning by CNNs in the training phase.

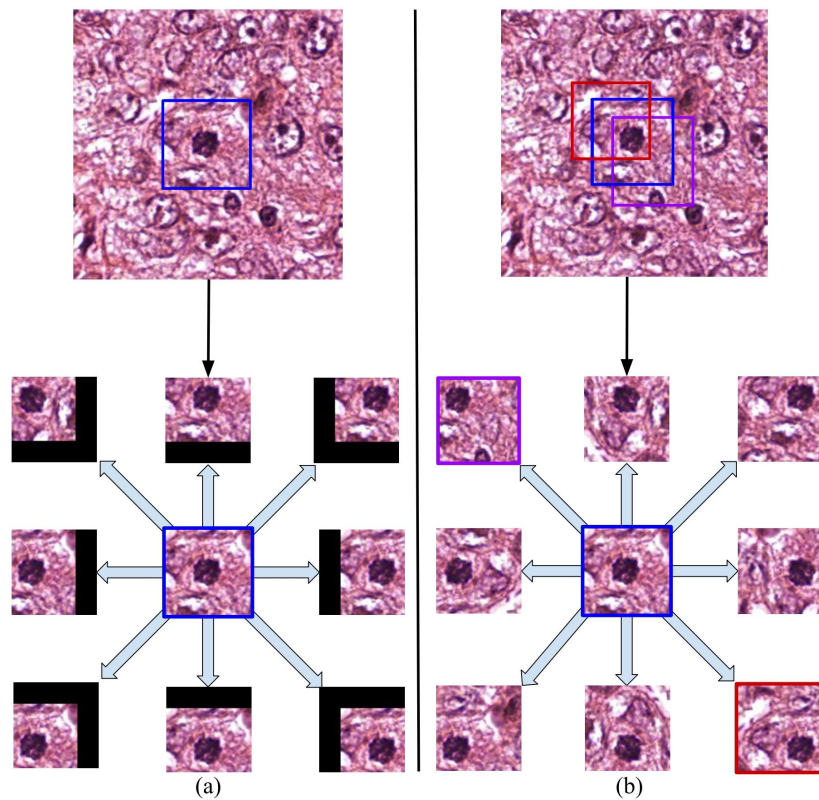


Figure 3.5: (a) Representation of conventional context non-preserving data augmentation (CNDA), (b) Context preserving data augmentation (CPDA) used in the proposed method.

Normally data augmentation techniques like translation, rotation, flipping etc. (Shorten and Khoshgoftaar, 2019), are used to counter shortage of training samples and class-imbalance. In conventional augmentation of the image patches, operations like translation, rotation etc., result in no-data regions typically having zero as pixel values. This leads to the loss of context information of the target object and disables an unfairly large number of neurons in the neural network to negatively impact the learning process. This can cause misclassification in testing since test samples are not augmented, and hence free from such no-data regions or zero pixels. In the proposed method, the extraction of the image patches is combined with a data augmentation technique referred to as context preserving data augmentation (CPDA). In the case of CPDA, cropping of image patches and data augmentation is combined in such a way that for a target object (mitotic cell), multiple patches are cropped. For each patch, the crop window is adjusted around the mitotic cell to create the effect of translation in multiple directions. Figure 3.5(a) demonstrates conventional translation based augmentation and Figure 3.5(b) shows the representation of context preserving augmentation. In this way, CPDA performs data augmentation for the mitotic image patches by preserving the context in the original image. The effect of CPDA on the performance of the method in comparison with the context non-preserving augmentation is discussed in the results section (Section 3.3.4).

### **3.2.4 CNN Training and Evaluation**

Deep learning, especially convolutional neural networks (Gu *et al.*, 2018), are widely used in pathology image analysis (Srinidhi *et al.*, 2020) these days. A deep convolutional neural network forms the backbone of the proposed method by classifying the patches containing tumor cells as mitotic or non-mitotic. Many of the recent CNNs are experimented to identify the one that performs better for this task. Result of this comparative study is given in Section 3.3.4. The outcome of the study indicates that DenseNet (Huang *et al.*, 2017) is the suitable architecture among the candidate CNNs

considered. DenseNet is a proven CNN architecture for image analysis tasks. It is made of multiple dense blocks in which each layer is connected to every subsequent layer within the dense block in a feed forward manner. Dense blocks are separated by transition layers consisting of convolution and max-pooling operations. DenseNet has several advantages such as resistance to vanishing gradient problem, less trainable parameters, low computational requirement etc., over its contemporary architectures. DenseNet architecture has three different configurations such as DenseNet121, DenseNet169, and DenseNet201. In the comparative study, it is observed that DenseNet121 gives the best results among these three configurations. Moreover, the number of trainable parameters is much less in DenseNet121, which significantly reduces the computations and hence the training time.

In the proposed method, DenseNet121 architecture is configured as a binary classifier. This architecture consists of 121 layers of trainable weights. The number of trainable parameters in this network are 5,245,568. For the intermediate layers ReLU is used as the activation function and Softmax is used for the final layer. Categorical cross entropy is the loss function used in this model. Adam optimizer with a learning rate of 0.0003 is used for weight optimization. The other hyperparameters used in the model are, batch size = 32, dropout = 0.3, and EPOCHs = 100. For finalizing the model hyperparameters, one random train-test division of MITOS-ATYPIA (ICPR, 2014) dataset is used.

Mitosis detection is posed as a binary classification problem in the proposed method. Every image patch is to be classified as mitotic or non-mitotic. An image patch is labeled as mitotic if there is at least one mitotic nucleus present in it and non-mitotic otherwise. Class imbalance between the two classes is addressed using data augmentation as described in the previous section. For training, every mitotic sample is augmented at a ratio of 1:20 using CPDA and other additional conventional techniques to match the total number of non-mitotic samples. For testing, a similar procedure is followed as in case of training except that data augmentation is not done on test image patches

to retain the class imbalance. Retention of the original class imbalance in HPF images while testing is essential to get a realistic performance measure of the model.

## 3.3 Experimental Results & Discussion

### 3.3.1 Dataset

The most commonly used dataset for mitosis detection is MITOS-ATYPIA (ICPR, 2014) released as part of the MITOS-ATYPIA grand challenge. This dataset contains labeled image data for two tasks related to breast cancer grading i.e., mitosis detection and nuclear atypia scoring. For mitosis detection, there are 2400 training images captured at  $40\times$  magnification using two scanner models Aperio Scanscope XT (1200 samples,  $1539 \times 1376$  pixels each) and Hamamatsu Nanozoomer 2.0-HT (1200 samples,  $1663 \times 1485$  pixels each). These images were captured randomly from pathology slides of breast cancer patients and analyzed by two expert pathologists to label the mitotic cells. Among the training images, 760 images have at least one mitotic cell present. Out of these, 80% of the images are utilized in training and the remaining for testing to decide the model hyperparameter values. From the training set, image patches of dimension  $96 \times 96$  pixels that contain the mitotic cells are extracted and augmented to match the number of non-mitotic image patches. The patch size is finalized based on the experiments on different patch sizes. A total of 44,180 image patches are created to train the neural network with equal share of mitotic (with augmentation) and non-mitotic (without augmentation) samples. Non-overlapping image patches are extracted by covering all the candidate nuclei detected in segmentation and used in testing. This has helped to further reduce the class imbalance between target classes, compared to patch extraction by centering every candidate nucleus.

There is another public dataset MITOS (ICPR, 2012) with 100 HPF images (70 for training & 30 for testing). Since the number of training images are less, this dataset



is not found to be suitable to train DenseNet independently. Instead, this dataset is used to supplement the MITOS-ATYPIA dataset and to increase the number of training and testing samples. Hence two sets of experiments are conducted, first one with MITOS-ATYPIA dataset and the second one with a combined dataset created by merging MITOS and MITOS-ATYPIA datasets.

### 3.3.2 Experiment Setup

The proposed method can be logically divided into two phases. One is the pre-processing phase that includes color normalization, candidate cell detection, and patch creation with data augmentation. This phase is performed using an Intel Xeon processor with 64 GB RAM and common python/matlab libraries. Second phase of the method is training and testing the deep CNN used i.e, DenseNet121. For this phase, a Tesla V100 GPU with 32 GB RAM and python framework Keras with Tensorflow as the backend are used.

Cross validation is considered as a preferred approach to validate the generalizability of a deep learning model when there is deficiency of data samples. For the proposed method, a 5-fold cross validation is adopted. The entire dataset is randomly divided into five disjoint sets of equal size. Five different train/test data folds are created using these disjoint sets such that each set is used exactly once as the test set and remaining sets combinedly as the training set. This process is equivalent to conducting the conventional training, validation, and testing five times with mutually exclusive random test sets of unseen data each time. This ensures that every sample in the dataset appears in the test set exactly once. As a result, 5-fold cross validation is a better way to assess the model and especially useful when the dataset is small with class imbalance present. All possible mitotic and non-mitotic patches of size  $96 \times 96$  pixels are extracted from the training set of every fold. Mitotic patches are augmented to match the number of non-mitotic patches using the techniques described before.

Cross validation experiments are carried out with three different compilations of the dataset while the process pipeline and CNN configuration are kept the same. The three dataset compilations are as follows.

*i) MITOS-ATYPIA with CPDA:* In this the mitotic samples are augmented by the proposed context preserving data augmentation (CPDA) as described in the methodology (Section 3.2.3).

*ii) Combined dataset with CPDA:* This dataset is formed by combining MITOS-ATYPIA and MITOS datasets. MITOS is a small dataset of 70 training images and 30 test images, suitable in conventional machine learning methods but not large enough to train advanced CNNs. In this compilation also the mitotic samples are augmented with CPDA.

*iii) Combined dataset without CPDA:* The mitotic samples in the combined dataset are augmented in the conventional context non-preserving manner (CNDA) as described in Section 3.2.3.

### **3.3.3 Evaluation Metrics**

Accuracy is not considered to be a suitable metric when there exists class imbalance in the test samples (Buda *et al.*, 2018; Chawla, 2009). In this binary classification problem, the positive class of mitotic cells are far less compared to non-mitotic cells and mimics. Hence, most of the existing methods for mitosis detection have used Precision, Recall, and F1 Score as the metrics for evaluation. For a fair comparison with the existing methods, the performance of the proposed method is captured using the same set of metrics. The definitions of these metrics are given in Appendix A.

Table 3.2: Result of the experiments carried out to choose the suitable CNN architecture for the proposed method.

<b>CNN Architecture</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
VGG16 (Simonyan and Zisserman, 2014)	48.99	70.19	57.71
ResNet50 (He <i>et al.</i> , 2016)	46.29	75.00	57.25
ResNet101 (He <i>et al.</i> , 2016)	55.56	55.29	55.42
NasNetLarge (Zoph <i>et al.</i> , 2018)	45.37	73.08	55.99
DenseNet169 (Huang <i>et al.</i> , 2017)	48.23	72.12	57.80
DenseNet201 (Huang <i>et al.</i> , 2017)	51.85	<b>80.77</b>	63.16
DenseNet121* (Huang <i>et al.</i> , 2017)	<b>73.05</b>	60.64	<b>66.27</b>

\*Used in the proposed method

### 3.3.4 Results

The results of the experiments conducted with the three different dataset compilations are presented here. Since a CNN forms the mainstay of the proposed method, selection of an appropriate CNN architecture was a crucial decision to make. For this purpose, experiments are carried out using many of the state-of-the-art CNN architectures. One random train-test split of the combined dataset was used in these experiments. Table 3.2 presents the results of the experiments for choosing the CNN. It was found that DenseNet architectures fared better in this comparative study. This has been the motivation for choosing a DenseNet architecture for the proposed method. Even though the DenseNet201 variant has given a higher recall value, it is computationally more expensive with nearly three times trainable parameters compared to DenseNet121. Hence, the DenseNet121 variant is chosen for the proposed method. If the additional computational complexity is ignored, DenseNet201 can also be a good choice for the CNN classifier to get better results.

Table 3.3 shows the results of 5-fold cross validation on the MITOS-ATYPIA dataset with the CPDA approach for augmentation. This gives mean values of 57.73, 60.94, and 59.29 for precision, recall, and F1 score respectively. Fold 1 is found to give the best results among the 5 folds. However, the average score of all folds is the reliable indicator of how well the model can perform on random unseen data.

Table 3.3: Result obtained using five-fold cross validation of base dataset MITOS-ATYPIA with CPDA.

<b>Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Fold 1	64.67	71.04	67.71
Fold 2	31.25	75.01	44.11
Fold 3	70.56	59.77	64.72
Fold 4	68.53	58.68	63.22
Fold 5	53.62	40.21	45.96
<b>Average</b>	<b>57.73</b>	<b>60.94</b>	<b>59.29</b>

Table 3.4: Result obtained using 5-fold cross validation of combined dataset with CPDA.

<b>Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Fold 1	73.05	60.64	66.27
Fold 2	32.01	88.88	47.06
Fold 3	70.34	57.51	63.28
Fold 4	60.25	69.23	64.42
Fold 5	57.14	52.17	54.54
<b>Average</b>	<b>58.56</b>	<b>65.69</b>	<b>61.91</b>

In the next set of experiments, the impact of supplementing the base dataset MITOS-ATYPIA with images from the MITOS dataset captured in a different setting is studied. Table 3.4 shows the results of the 5-fold cross validation of this combined dataset. Average values of precision, recall, and F1 score are 58.56, 65.69, and 61.91 respectively. Compared to the previous results of the MITOS-ATYPIA dataset, here the results show improvements for all the three measures. This shows that the proposed method continues to give improved performance as the dataset gets bigger with more samples, even when the additional samples are from a different dataset altogether. It also indicates that the performance of the model can improve further if more training samples are added. In that way, the proposed method offers a general framework for mitosis detection that is more resilient to dataset variations resulting from staining differences and acquisition setting.

In Figure 3.6, the learning pattern of DenseNet121 is shown graphically. Training loss, training accuracy, validation loss, and validation accuracy are plotted against the training EPOCHs for the base dataset (Figure 3.6(a)) and the combined dataset (Fig-

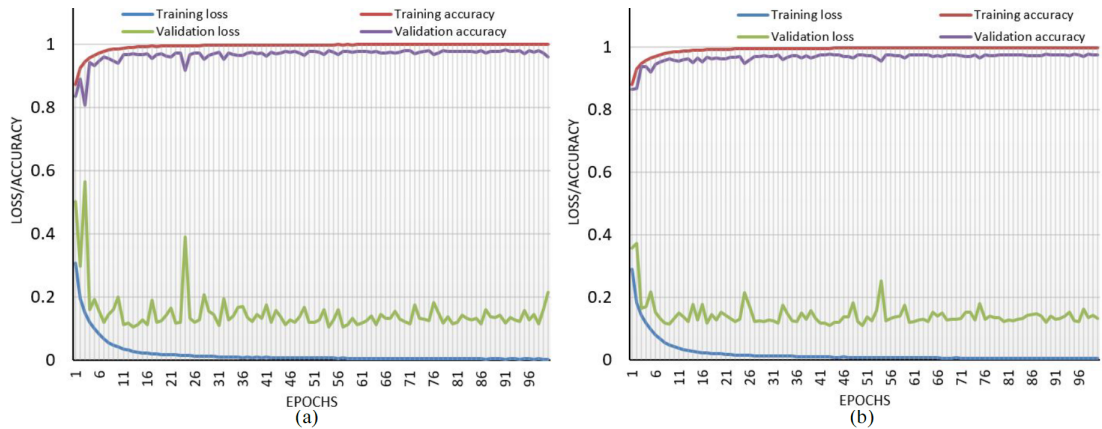


Figure 3.6: Learning pattern of the CNN using (a) base dataset and (b) combined dataset.

ure 3.6(b)). It is noticeable from the validation curves that the learning is smoother (less fluctuations) in case of combined dataset especially in the second half of training. This indicates improved stability of the model trained with the combined dataset. The effectiveness of a classification system at various thresholds is normally captured using receiver operating characteristic (ROC) curves and area under ROC curves (AUC) for each class. However, in classification problems with large class imbalance between positive and negative classes ROC curves give an over-optimistic representation of the performance. In such cases precision-recall plots give a realistic representation of the classification performance for each class (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006). The severe class imbalance between mitotic and non-mitotic cells has been the motivation to consider precision-recall curves over ROC curves. Figure 3.7 shows the precision-recall curves obtained for all the five folds of cross-validation. The average precision (AP) obtained for all thresholds is shown for each class. These plots show a balancing effect of precision and recall for the positive class of mitotic cells and are in line with the results shown in Table 3.4. For the negative class of non-mitotic figures, the effectiveness of the system is much better even though its significance in this problem is low.

Confusion matrix gives a detailed picture of the classification with respect to the four possible outcomes for input samples such as true positive (TP), true negative (TN),

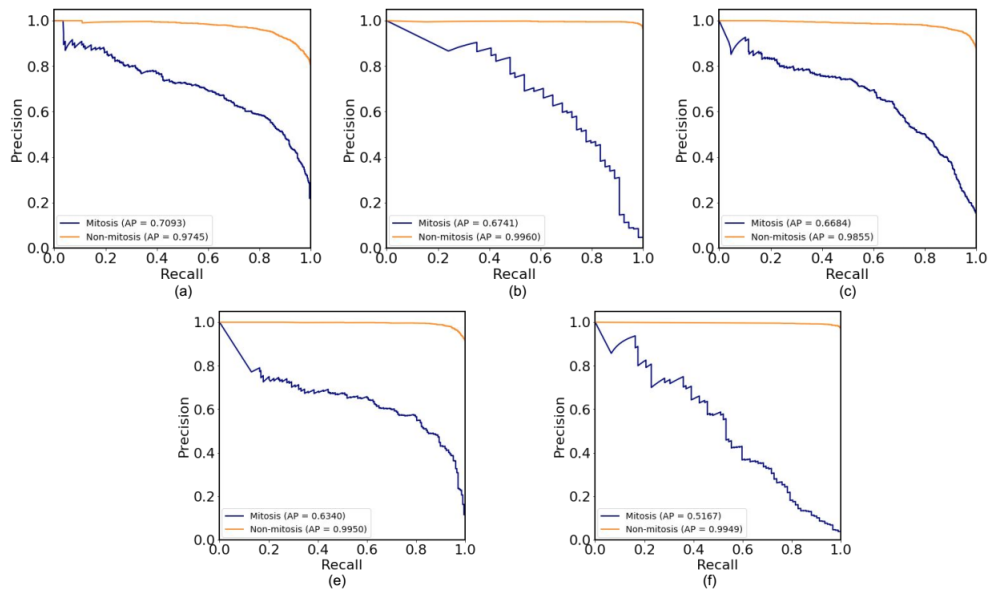


Figure 3.7: Precision-Recall curve and average precision (AP) obtained for the mitotic and non-mitotic cell classification using different folds of the combined dataset (a) Fold 1, (b) Fold 2, (c) Fold 3, (d) Fold 4, and (e) Fold 5.

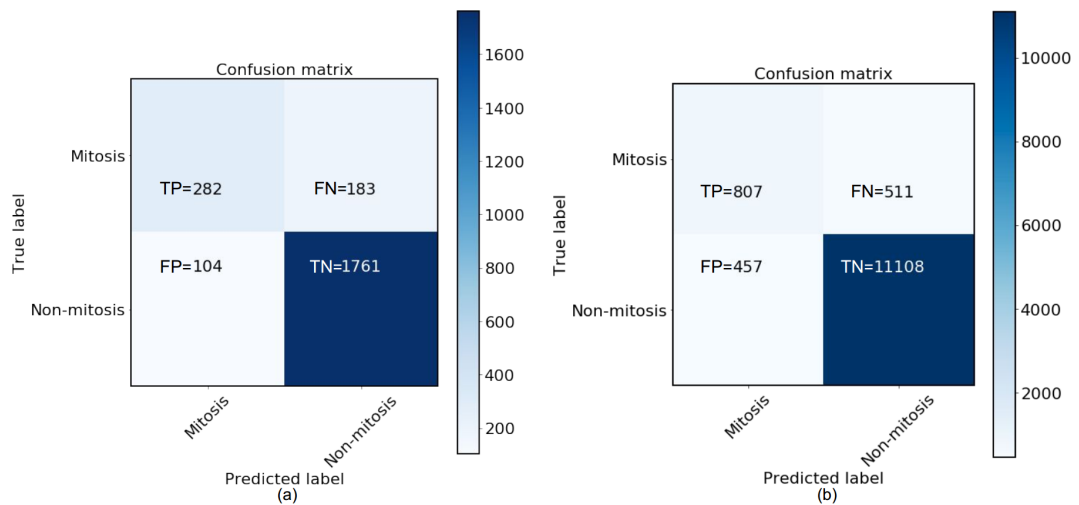


Figure 3.8: Representative confusion matrices obtained for the combination of datasets MITOS and MITOS-ATYPIA (a) Fold 1, (b) All folds combined. Mitosis is the positive class and non-mitosis figures constitute the negative class in this binary classification problem.

false positive (FP), and false negative (FN) to which each test sample is mapped. In Figure 3.8, confusion matrix obtained for the fold 1 (Figure 3.8(a)) and the combined matrix for all five folds are presented (Figure 3.8(b)). In both cases, the model is effective

Table 3.5: Comparison of context preserving data augmentation (CPDA) and conventional context non-preserving data augmentation (CNDA) in each fold of 5-fold cross validation.

CV-Fold	Augmentation	Precision	Recall	F1 Score
Fold 1	CNDA	76.47	50.32	60.70
	CPDA	73.05	<b>60.64</b>	<b>66.27</b>
Fold 2	CNDA	23.52	74.07	35.71
	CPDA	<b>32.01</b>	<b>88.88</b>	<b>47.06</b>
Fold 3	CNDA	65.81	51.70	57.91
	CPDA	<b>70.34</b>	<b>57.51</b>	<b>63.28</b>
Fold 4	CNDA	63.06	53.36	57.81
	CPDA	60.25	<b>69.23</b>	<b>64.42</b>
Fold 5	CNDA	58.88	32.60	41.95
	CPDA	57.14	<b>52.17</b>	<b>54.54</b>
Average	CNDA	57.55	52.41	54.86
	CPDA	<b>58.56</b>	<b>65.69</b>	<b>61.91</b>

in classifying the negative class of non-mitotic cells as indicated by the high value of TN and relatively low value of FP. However, the effectiveness of mitotic cell classification is moderate with a high proportion of FNs. These figures suggest the challenging nature of the mitotic detection problem. Even though the strategies like combining datasets and augmentation yielded positive results, the need for further research is visible from these outcomes.

The impact of context preserving data augmentation technique applied in the proposed method is studied in a separate set of experiments. Table 3.5 presents the comparison of CPDA approach with the CNDA using 5-fold cross validation on the combined dataset. The dataset folds and experimental setup were kept the same for the two sets of experiments, one with CPDA and other with CNDA. It is observed that the CPDA approach clearly gives superior performance over CNDA. Recall values and F1 scores are consistently better for all the folds in cross-validation when CPDA is used. Average values of precision, recall, and F1 score using CPDA show significant improvement over the corresponding values using CNDA.

The result of the proposed method is compared with the deep learning methods that

Table 3.6: Comparison of the proposed method with the state-of-the-art deep learning methods based on MITOS-ATYPIA dataset. (A: Aperio scanner images, H: Hamamatsu scanner images, M: MITOS dataset).

Method	Precision	Recall	F1 Score
Deep ResNet (Li <i>et al.</i> , 2018)	43.10	44.30	43.70
Deep Cascade Network (Chen <i>et al.</i> , 2016a)	41.10	47.80	43.70
RCNN (Cai <i>et al.</i> , 2019)	53.00	66.00	59.50
DCNN + Wavelets (A) (Das and Dutta, 2019)	54.40	57.60	55.90
DCNN + Wavelets (H) (Das and Dutta, 2019)	57.40	62.20	59.70
DCNN + Wavelets (A-H Avg.) (Das and Dutta, 2019)	55.94	59.94	57.87
<b>Proposed method (A&amp;H)</b>	57.73	60.94	59.29
<b>Proposed method (A&amp;H&amp;M)</b>	58.56	65.69	61.91

used the MITOS-ATYPIA dataset for a fair comparison. The compared methods have used either a single train-test split of the dataset (Li *et al.*, 2018; Chen *et al.*, 2016a) or cross validation (Cai *et al.*, 2019; Das and Dutta, 2019). Evaluation of the model with a single train-test split may not reflect a realistic performance since only a fraction of the dataset is used for testing. Possibility of biased results is high in this case due to over-fitting or a bias in the test sample selection. The five-fold cross validation carried out is equivalent to the creation of five train-test splits and evaluation of the model with each of them. This approach makes sure that every sample in the dataset appears once as a test sample in any one of the folds. It also eliminates the possibility of biased results due to over-fitting. The final result is computed by averaging the results of all the folds. Comparison of the proposed method with the other deep learning-based methods is given in Table 3.6. It can be seen from the table that the proposed method gives better performance over the existing methods on the base dataset MITOS-ATYPIA. The values of metrics *Precision*, *Recall*, and *F1Score* are improved by 3.2%, 1.7%, and 2.5% respectively over the state-of-the-art method. Using the combined dataset, the results show further improvement of 4.7%, 9.6%, and 6.9%.



### 3.3.5 Discussion

There are various factors which make automated mitosis detection a challenging task. The major ones are varying shape and size of mitotic nuclei, their similarity with apoptotic cells, class imbalance between mitotic and non-mitotic samples in HPF images, staining variations in slide preparation, limited size of the datasets etc. Here, the impact of the two strategies applied to overcome the dataset size limitation and class imbalance are discussed.

#### 3.3.5.1 Impact of Combining Datasets

Most of the methods in literature have used shallow CNN architectures (Chen *et al.*, 2016a; Cai *et al.*, 2019; Das and Dutta, 2019) due to the limited dataset size and class imbalance. These problems have been bottlenecks for using very deep CNNs in mitosis detection due to poor learning by the networks. Moreover, all existing methods in the literature are independently trained on each dataset and tested on unseen samples from the same dataset. Such a trained model is less likely to give similar performance on test samples from a new dataset and less useful for implementation in clinical practice. A recommended model would be the one which is more resilient to variations across datasets resulted by slide preparation, image acquisition setup etc. Instead of training independent models for each new dataset, a model that continues to improve the performance in sync with constant addition of new training samples to a single training pool is more suitable in practice. The proposed method is first of its kind to adopt this paradigm for mitosis detection. Color normalization (Reinhard *et al.*, 2001) in the first stage brings the images from different datasets to a common color level to alleviate the variations related to staining and acquisition. An advanced deep learning architecture capable of continued learning from new data samples introduced in the dataset is chosen. The experimental results testify this. The base dataset (ICPR, 2014) has 760 HPF images with at least one mitotic cell in each. The result obtained using this dataset is given in Table 3.3. Adding another 100 images from (ICPR, 2012) to this base dataset

gives notable improvement in the results as shown in Table 3.4. This improvement points to the prospects of combining smaller datasets with appropriate normalization techniques to create a large dataset to meet the data requirement of deep neural networks. Methodologies like deep learning and research problems that require a large dataset may consider this approach to create sufficient data to train the algorithms. Such methods will be more generalizable and better accommodative to unseen test data as required in clinical application. In mitosis detection and many related pathology image analysis tasks, the possibility of creating a large dataset from a single facility is remote, considering the manual effort required from pathologists to acquire and annotate slide images.

### **3.3.5.2 Impact of Context Preserving Augmentation**

Class imbalance is a serious hurdle for adoption of deep learning approaches in many image analysis tasks (Johnson and Khoshgoftaar, 2019; Wang and Yao, 2012). It is the skewed distribution of target class samples in the dataset, leading to poor or biased learning. In the case of mitosis detection, the class imbalance between the positive class (mitosis) samples and negative class (non-mitosis) samples is huge. In an HPF image at  $40\times$  magnification typically there are 0 – 5 mitotic cells whereas non-mitotic cells may go up to a few hundreds. This severe class imbalance necessitated heavy augmentation of the mitotic samples to match the number of non-mitotic samples present in large numbers. The context preserving image data augmentation (CPDA) applied in this work made a positive impact on the performance. The conventional augmentation, based on geometric transformations (Shorten and Khoshgoftaar, 2019), leads to loss of original context of the mitotic nuclei in the augmented patches and turn-off a large number of neurons in the neural network in the training stage. As a result of CPDA, the contextual information of the mitotic nuclei is preserved in the augmentation process. Results in Table 3.5 shows the impact of CPDA compared to the conventional augmentation that does not care about preserving the context of the target object. However, CPDA in

its original form is suitable only for patch-based approaches where the actual training images are extracted patches from images of high dimensions. Many of the pathology image analysis tasks follow the pattern of patch-based processing.

### **3.4 Summary**

In this chapter, a new deep learning-based method for automated mitosis detection in H & E-stained histopathology images of breast cancer is proposed. The method involves multiple stages such as color normalization of slide images, detection of candidate cells, and patch-based training of an advanced CNN to classify mitotic and non-mitotic cells. Class imbalance between the two target classes necessitated heavy augmentation of mitotic samples to match the number of non-mitotic samples to train the CNN. The augmentation is carefully done in a context preserving way to yield improved results. The base dataset is extended by merging another dataset acquired in a different setup altogether to address the data insufficiency. The proposed method shows adaptability to the additional dataset by giving better performance compared to the base dataset. This shows that the model continues to improve from the new training samples in spite of the variations in the images acquired from different settings. The effort to apply an advanced CNN like DenseNet on a skewed dataset for mitosis detection has shown encouraging results. The scope for future improvements to the proposed method is by applying strategies such as enhancing the dataset further, compensating for the class imbalance in the model design, and use of better CNNs.



## CHAPTER 4

# AUTOMATED NUCLEAR ATYPIA SCORING OF BREAST CANCER

In this chapter, a novel deep learning-based framework for automated nuclear atypia scoring of breast cancer is proposed. The framework consists of three major phases namely preprocessing, deep learning, and postprocessing. The original three-class problem of slide level atypia scoring is reformulated as a six-class problem of nuclei classification to enable the effective use of deep learning algorithms to address this task. Subsequently, a deep convolutional neural network (CNN) is used to classify the six classes of nuclei present in slide images. The output of this classifier is processed to predict the nuclear atypia score of the input slide image. The publicly available slide image dataset MITOS-ATYPIA is used for the experiments. The proposed method gives a performance that exceeds the state-of-the-art by a significant margin with the results 0.8766, 0.8760, and 0.8745 for the metrics precision, recall, and F1 score respectively. The improvements in these metric values by 13.93%, 9.89%, and 11.90% over state-of-the-art method vindicate the effectiveness of the proposed framework in automated atypia scoring of breast carcinoma.

### 4.1 Introduction

Nuclear atypia (also known as nuclear pleomorphism) is an integral factor of breast cancer grading as per Nottingham Grading System (NGS) (Elston and Ellis, 2002). Nuclear atypia refers to the degree of morphological distinction of malignant tumor nuclei from normal nuclei. Compared to the parent tissue cells which are generally uniform in appearance, tumor cells show large variations in their features such as size, shape, number of nucleoli, and chromatin distribution (Stierer *et al.*, 1991; Kristiansen, 2018). These variations are valuable indicators of the aggressiveness of cancer and hence the assessment of these variations forms an important parameter of breast cancer grading and prognostication (Pienta and Coffey, 1991). Unlike the mitosis count parameter,

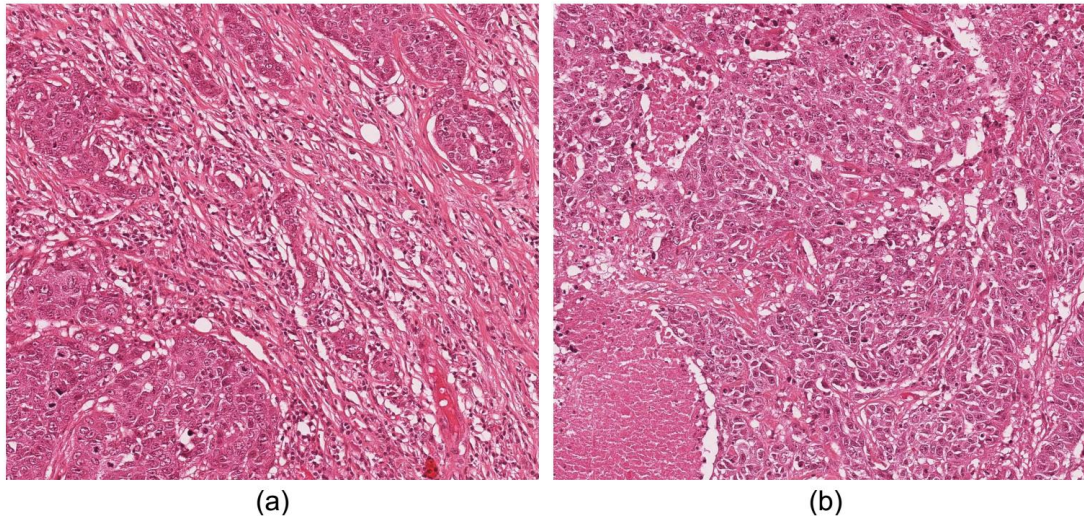


Figure 4.1: Sample slide images to demonstrate complexity and structural diversity within histopathology slide images. Major portions of the slide images are occupied by (a) stroma, tumor cells, and lymphocytes, (b) tumor cells, necrosis, stroma, and fat globules. A closer view of different regions is given in Figure 4.8.

which is objective and has well-defined criteria, nuclear atypia scoring is largely subjective (Dunne and Going, 2001) mainly due to practical difficulty in measuring multiple contributing factors in a uniform way. This subjectivity makes manual atypia scoring prone to errors, intra- and inter-observer variations, and low reproducibility (Frierson Jr *et al.*, 1995). Fully automated or computer-assisted atypia scoring can solve these problems associated with manual scoring (Gandomkar *et al.*, 2019; Das *et al.*, 2020b).

#### **4.1.1 Challenges in Automated Nuclear Atypia Scoring**

H & E-stained histopathology slide analysis is the universally accepted and cost-effective procedure for nuclear atypia scoring and breast cancer grading. When it comes to the automation of this procedure through image analysis, there are several challenges existing. The major challenges found from the literature and experienced during the development of this work are discussed in this section.

#### **4.1.1.1 Complexity of Histological Slide Images**

Conventionally nuclear atypia scoring is performed at  $20\times$  magnification of the histology slide, representing a relatively larger field-of-view through the microscope. The slide images captured at this magnification have an amalgamation of diverse structures such as malignant tumor cells, lymphocytes, stroma cells, necrotic cells, hemorrhages, lipids, etc. However, only the malignant tumor regions and morphological attributes of tumor cells are considered for atypia scoring. In fully automated methods, the other cellular structures are irrelevant and often lead to performance degradation since an accurate delineation of the tumor regions from the rest is difficult. Figure 4.1 shows sample slide images from two different subsets of the public dataset MITOS-ATYPIA (ICPR, 2014) used in most of the automated atypia scoring methods in the literature. Most parts of the images are occupied by tumor cells, lymphocytes, stroma (connective tissues), necrosis, and fat globules. Multiple clusters of these components are spread across the image unevenly to make the overall structure highly complex.

#### **4.1.1.2 Inter-class Similarity and Intra-class Variations**

Deep learning algorithms require large amounts of training data. This is because the discriminative features of the target classes are learned from the labeled training samples presented to the algorithm, and more training samples lead to improved learning by the model. The MITOS-ATYPIA dataset contains slide images from two different slide scanners namely Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. Images from these two scanners vary in their color intensity levels due to staining variations and make/configuration of the scanners. In many cases, similarity between slide images that belong to different atypia score classes is high as evident from the level of pathologists' disagreement in atypia scoring (Dunne and Going, 2001). Figure 4.2 demonstrates instances of inter-class similarity present in the slide image dataset used for nuclear scoring. Figure 4.2(a) and Figure 4.2(b) are extracted samples of tumor regions from image subset A10. Even though visually both the images look very similar

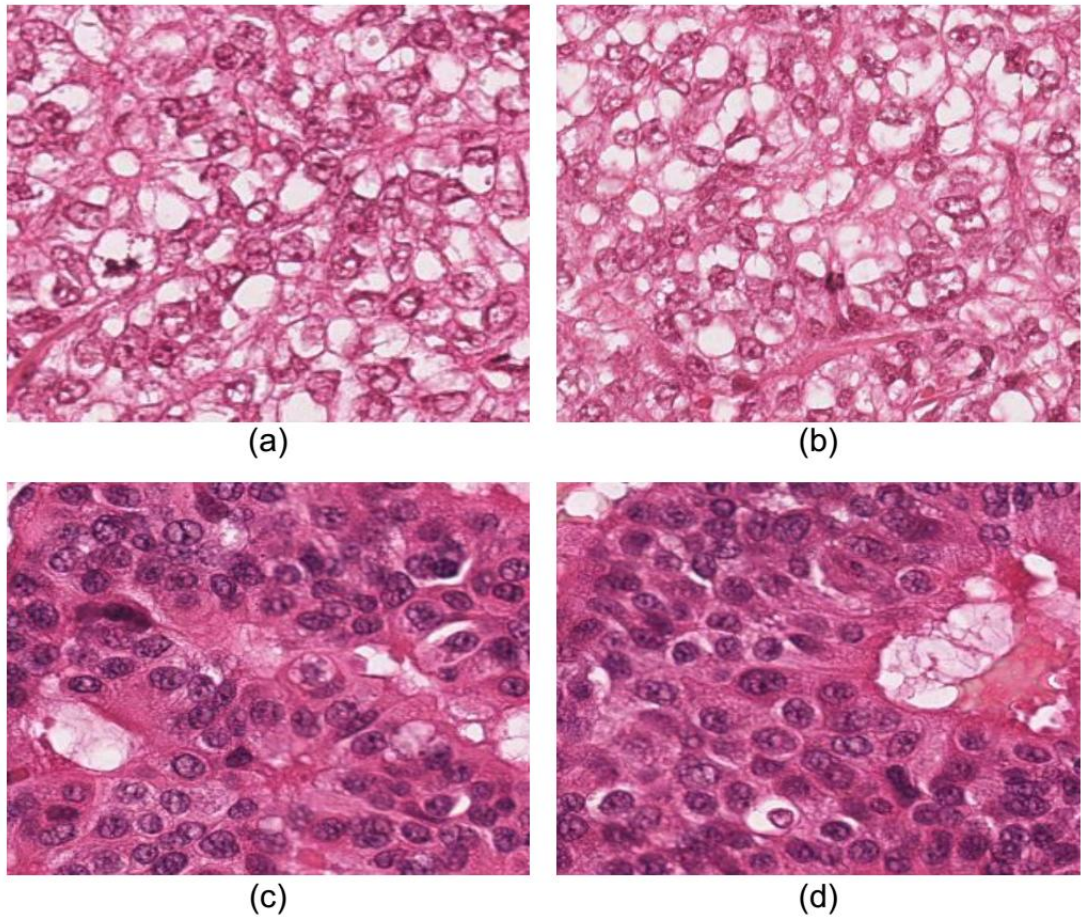


Figure 4.2: Inter-class similarity of slide images. From subset A10: (a) Score 2 slide image, (b) Score 3 slide image; From subset A11: (c) Score 2 slide image, (d) Score 3 slide image.

they are labeled as score 2 and score 3 respectively. In the same way Figure 4.2(c) and Figure 4.2(d) are samples from subset A11 and labeled as scores 2 and 3. These two samples also look very much identical but annotated with different atypia scores. Such scenarios are extremely challenging in both manual and automated atypia scoring.

Intra-class variation is another problem in automated atypia scoring. The appearance of the malignant tumor images of the same atypia score may have large variations. Figure 4.3 illustrates intra-class variations present in the tumor regions of slide images from different subsets of the MITOS-ATYPIA dataset. All the samples in this figure are labeled with nuclear atypia score 2. But the variations in color intensity, mor-



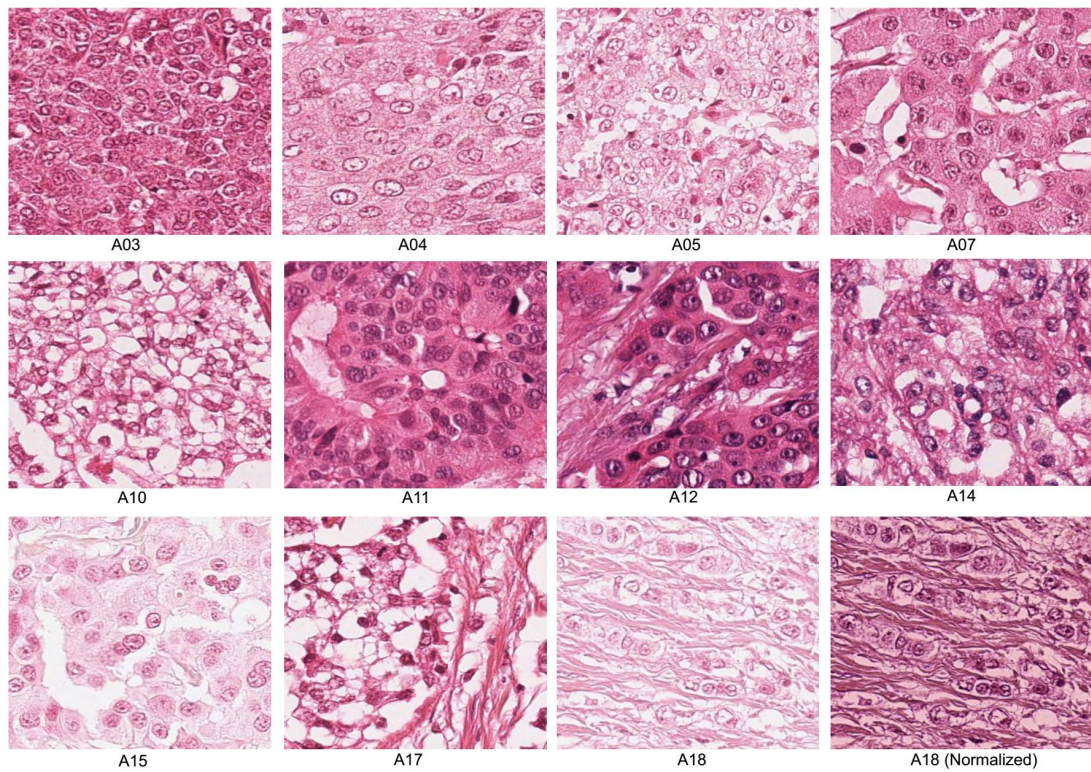


Figure 4.3: Intra-class variations in score 2 type slide images from different subsets (A03, A04 etc.) of the MITOS-ATYPIA dataset. These samples vary substantially in appearance even though they all have the same atypia score of 2.

phology of tumor cells, texture, etc., are apparent to even a non-expert observer. Most of the existing methods for automated nuclear atypia scoring are based on the hand-crafted features extracted from these images. It is challenging to design a generalized model based on handcrafted features to capture these levels of intra-class variations and classify the unseen image samples accurately. This may be the reason behind the poor performance of handcrafted feature-based methods in the literature. In general, these challenges apply to other similar histopathology image analysis tasks as well. In the proposed framework these challenges are addressed by reformulating the problem and with the effective use of deep learning.

Apart from a few earlier methods, significant research on nuclear atypia scoring happened after the grand challenge MITOS-ATYPIA (2014), which released a large public dataset of 600 slide images of  $20\times$  magnification for this task. Since then, several

methods have been published for solving this clinically relevant problem. However, the existing methods have not achieved the performance level required for application in clinical practice. This can be partly attributed to the challenging nature of the problem. It is also observed that the potential of advanced deep learning algorithms is not fully exploited to address this problem. In its original form, the dataset is not suitable to apply high performing deep learning algorithms like CNNs. The reasons for this are the high dimension of the slide images (Aperio:  $1539 \times 1376$ , Hamamatsu:  $1663 \times 1485$ ), the structural complexity of the images, and the availability of only image-level atypia score in the dataset as ground truth. Computationally it is not feasible to feed such high dimensional images to CNNs directly. Even if it was feasible, the structural diversity within the images cannot be discriminatively learned by the CNNs with the limited number of slide images per scanner, after reserving a fraction for testing.

Nuclear atypia scoring is a three-class problem where each slide image is assigned with an atypia score of 1, 2, or 3. In the proposed framework, this is reformulated as a six-class problem through additional labeling and preprocessing of the training slide images. Six major nuclei classes are defined to categorize all nuclei present in slide images. This includes three-classes of cancerous nuclei that are used for nuclear atypia scoring (*scoring classes*) and another three classes that are not involved in the scoring process (*elimination classes*). In the absence of these elimination classes, nuclei of such types might end up being classified into any one of the scoring classes and affect the accuracy of scoring. A set of preprocessing steps are used to transform the dataset in a manner that becomes suitable to train an advanced CNN to classify all nuclei into appropriate classes. In the post-processing stage, the nuclei that are classified into scoring classes are used for atypia score prediction. Rest of the nuclei classified into elimination classes are ignored since they are not required for atypia scoring as per the pathology procedure.

The major contributions of this chapter are as follows:

- A novel framework is proposed for automated nuclear atypia scoring of breast cancer that closely resembles the pathologists' way of manual nuclear atypia scor-

ing. The framework has a futuristic design that provides flexibility to plug newer and better algorithms in its major phases such as preprocessing, deep learning, and post-processing to further improve atypia scoring performance.

- The intrinsic three-class problem of slide level nuclear atypia scoring is reformulated as a six-class problem of nuclei classification in the deep learning phase of the framework. This reduces complexity and makes it appropriate for the deep learning algorithm to learn the features that accurately discriminate different types of nuclei.
- The method proposed in this chapter follows this framework. The result obtained for this method is significantly improved over the state-of-the-art methods for automated atypia scoring. This performance testifies to the effectiveness of the proposed approach.

The rest of the chapter is organized as follows. The proposed framework and the method based on this framework are explained in Section 4.2. Experimental setup, results, and related discussions are presented in Section 4.3. Finally, the chapter is concluded with a summary of the proposed method and the outlook on the clinical usage of automated nuclear atypia scoring.

## 4.2 Proposed Framework

In the three-class problem of nuclear atypia scoring, a slide image from the tumor region is assigned a score of 1, 2, or 3 by a pathologist based on the features of malignant nuclei present. However, there are other types of non-malignant cells/nuclei present in slide images that are ignored by pathologists in the manual procedure. The proposed framework adopts the approach of segregating these two categories of nuclei (malignant & non-malignant) and using only malignant nuclei for atypia scoring. To achieve this objective, the problem is formulated as a six-class nuclei classification problem where the first three classes correspond to nuclei of score 1, 2, and 3. These classes are together referred to as *scoring classes*. The remaining three classes are nuclei of lymphocytes, necrotic cells, and stroma cells which are non-malignant. These classes are referred to as *elimination classes* since these are not involved in atypia scoring.

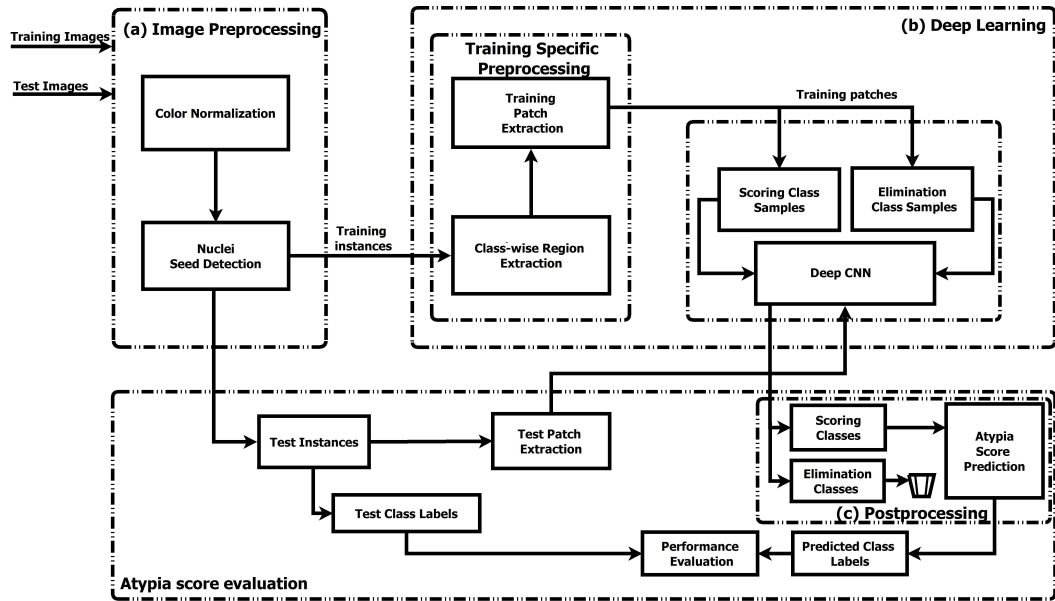


Figure 4.4: Overview of the proposed framework for automated nuclear atypia scoring. Three major phases in the framework are: (a) Image preprocessing, (b) Deep learning, (c) Post-processing. (Post-processing is involved only in the slide level evaluation of nuclear atypia score.)

The proposed framework for nuclear atypia scoring is outlined in Figure 4.4. This framework consists of three major phases namely i) Image preprocessing, ii) Deep learning, and iii) Post-processing. The image preprocessing phase is common to both training and evaluation. In this phase, color normalization and nuclei detection are performed on the input slide images. Color normalization is aimed at mitigating the color variations introduced in the slide images due to the staining differences or acquisition setup. Further, all the nuclei present in the slide images are detected. The deep learning phase of the framework starts with further preprocessing specific to training of the deep CNN used. Here, the three-class problem of nuclear atypia scoring is transformed into a six-class nuclei classification problem. This is done through class-wise extraction of nuclei regions from the training slide images. This process is explained in Section 4.2.2. The framework offers flexibility to use any suitable CNN classifier algorithm in the deep learning phase. The chosen CNN is trained to accurately classify the six classes of nuclei. For the prediction of slide level atypia score, all the nuclei detected in any given test slide are extracted as fixed size image patches and fed to the

trained CNN model. In the post-processing phase, nuclear patches that are classified into scoring classes are used to predict the atypia score corresponding to the input slide image. All the nuclei that are classified into any of the elimination classes are ignored as in the case of manual atypia scoring. Overall, the framework is designed in a generic manner that in each phase such as preprocessing, deep learning, and post-processing, it is possible to choose any existing or new algorithms suitable for the task. A concrete method pipeline is implemented based on this framework (the proposed method). In the following sections, the proposed method is explained in detail.

#### **4.2.1 Image Preprocessing**

Color normalization and nuclei seed detection are the major steps in the preprocessing phase of the proposed framework. There are several specialized algorithms available in literature to perform these tasks for H & E-stained histopathology images. In this section the algorithms applied in the preprocessing phase of the proposed framework are briefly discussed.

##### *Color normalization of slide images*

One of the challenges in histopathology image analysis is the color variations in the slide images. This is mainly due the variation in concentration of stains used and the configuration of scanners that are used for digitizing the slides. Tissue staining is a manual process and when done by different people, there is a possibility of stain variation causing color changes in the slide images. Such variations are counter-productive in CNN-like algorithms which tend to learn pathologically irrelevant color features. Color normalization is a common preprocessing step used to solve this problem (Onder *et al.*, 2014). In the MITOS-ATYPIA dataset, there are slide images from two different makes of the scanners, and images from these scanners show substantial color variation necessitating color normalization. There are several color-normalization methods available for histopathology images in literature (Reinhard *et al.*, 2001; Khan *et al.*,

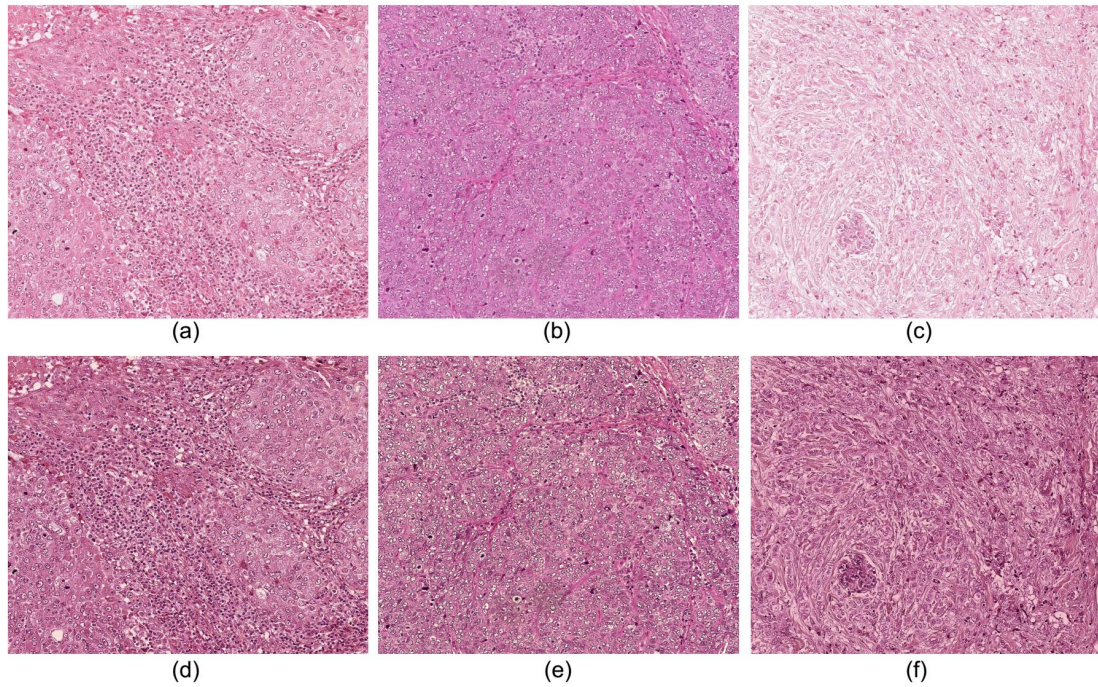


Figure 4.5: Color normalization of slide images. (a, b, and c) are the unnormalized images and (d, e, and f) are the corresponding normalized images.

2014; Li and Plataniotis, 2015; Vahadane *et al.*, 2016). The classic color normalization technique proposed by Reinhard *et al.* (2001) based on  $l\alpha\beta$  color space is applied in the proposed method. The method computes mean and standard deviation of a reference H & E image in  $l\alpha\beta$  space and matches the color statistics of the source image to this. Figure 4.5 shows the sample slide images from the MITOS-ATYPIA dataset which are captured using Aperio and Hamamatsu scanners. The images a, b, and c show visible color variations, which distracts the deep learning algorithms that are expected to learn morphological and textural features for atypia scoring. The images d, e, and f in Figure 4.5 show the corresponding normalized versions of a, b, and c. Color normalization is observed to perform contrast enhancement of over-exposed images as in the case of Figure 4.5(c). Normalized image Figure 4.5(f) has better contrast and visible texture than the original image.

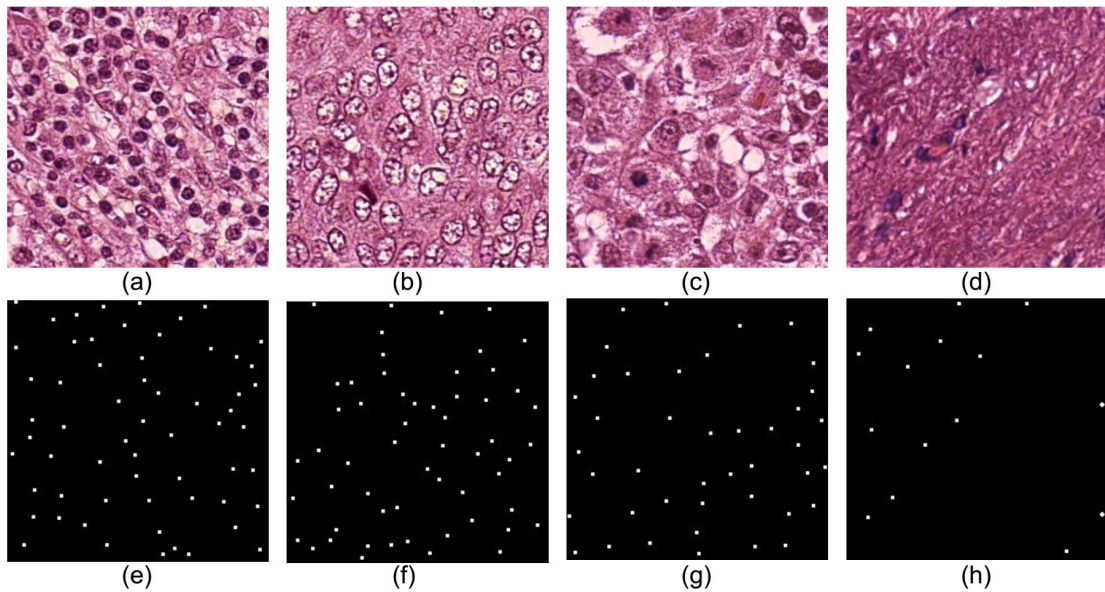


Figure 4.6: Nuclei seed detection in slide regions of different nuclei classes. (a, b, c, and d) are the slide regions of lymphocytes, score 2, score 3, and necrosis classes respectively, (e, f, g, and h) are corresponding nuclei detected.

#### *Nuclei seed detection*

Once the slide images are color-normalized, the next preprocessing step is to detect all the nuclei in the slide images. The nuclei detection and segmentation algorithm by Al-Kofahi *et al.* (2009) is used to detect the nuclei seed points. In this method, color deconvolution is used to separate hematoxylin and eosin components from the slide images. In H & E staining, hematoxylin attaches with the nucleoli to give them dark-purple color whereas eosin binds with cytoplasm and surrounding structures to give shades of pink. Color deconvolution helps to focus on nuclei spots in the image. Using graph-cut based binarization (Boykov and Kolmogorov, 2004b), the foreground is extracted from the hematoxylin component image (nuclear channel). Laplacian-of-Gaussian filtering with distance map-based adaptive scale selection is used to detect the nuclei seed points. These seed points are used to segment the nuclei using region adjacency graph coloring. In the proposed method, segmenting the nuclei has little relevance since only the nuclei seed points are required to extract image patches centered on the nuclei. Figure 4.6 shows the results of applying nuclei seed detection on slide im-

age extracts of different classes. This algorithm successfully detects all types of nuclei present in the slide image.

## 4.2.2 Deep Learning Stage

The deep learning phase of the proposed framework starts with a preprocessing step specific to the training of CNN. This step implements nuclear atypia scoring as a six-class problem. It involves class-wise region extraction for all the six classes of nuclei followed by nuclear patch extraction. A slide image captured at  $20\times$  normally contains several hundreds of nuclei. Class-wise region extraction helps to avoid the need to label every nucleus in the slide image individually with an appropriate class label for training the CNN. The fixed size nuclear image patches extracted from class regions are used to train the chosen deep CNN. Training specific preprocessing and CNN training are shown in Figure 4.7 and explained below.

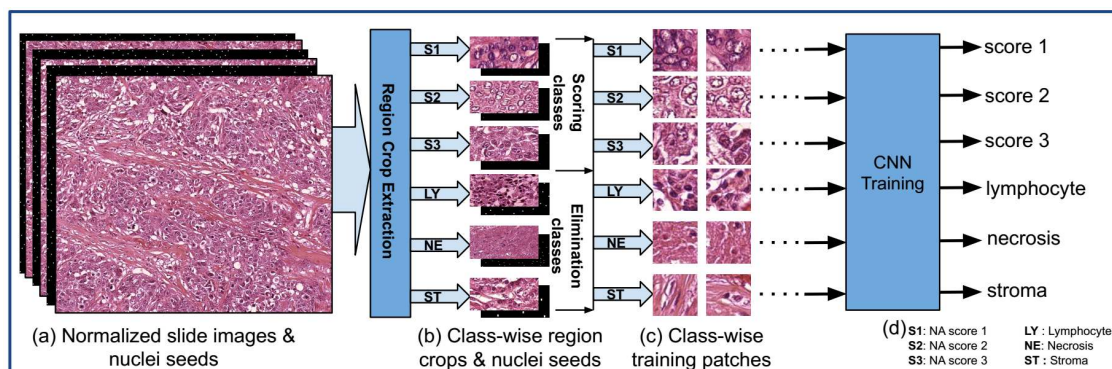


Figure 4.7: Training specific preprocessing and CNN training in the deep learning phase of the proposed framework.

### *Class-wise region extraction & training patch creation*

The most impactful factor behind the performance of the proposed framework is the six-class model designed for nuclear image patch classification. Atypia scoring is based on



the malignant tumor nuclei that are classified into three classes namely score 1, score 2, and score 3 according to their morphological features. However, a slide image typically contains several different types of non-malignant cells as well. These cells include lymphocytes, stroma cells, necrosis, etc., which are ignored by the pathologist in the routine atypia scoring procedure. In a three-class model, nuclei of such cells can get classified into one of these three classes and adversely affect the scoring accuracy. In deep learning-based methods, this problem is more prominent as the algorithms tend to classify any nuclei into one of the three classes. To counter this problem, the six-class classifier is designed where the first three classes are referred to as the *scoring classes* (SC) which include the three types of malignant tumor nuclei of scores 1, 2, and 3. The remaining three classes referred to as *elimination classes* (EC) are lymphocytes, stroma cells, and necrotic cells that form the major population of non-malignant cells/nuclei present in slide images. Figure 4.8 shows the sample region crops from slide images for all the six classes. Apparently, scoring class nuclei have clear discriminative features from the elimination classes. Hence a six-class model can potentially segregate the elimination class nuclei and the scoring class nuclei. Any other unlisted categories of non-malignant nuclei are more likely to be classified into one of the elimination classes rather than scoring classes due to their close similarity. This will not impact atypia scoring since elimination classes are ignored in the final scoring process.

The major challenge in formulating nuclear atypia scoring as a classification problem with more than three classes is that the MITOS-ATYPIA dataset has only an image-level score label (1, 2, or 3) for every slide image. This limitation is overcome by adding six new class-wise region labels in the training images (3 SCs + 3 ECs). Under the supervision of a senior pathologist, regions corresponding to all the six classes of nuclei are marked in the training images. Since nuclei of every class are normally present in clusters, region marking is fairly effortless and one time activity required only for training set. This way of region marking saved the mammoth task of labeling every nucleus in the slide image into one of the six identified classes. The marked regions in the training images are extracted for further processing. The region crops may vary

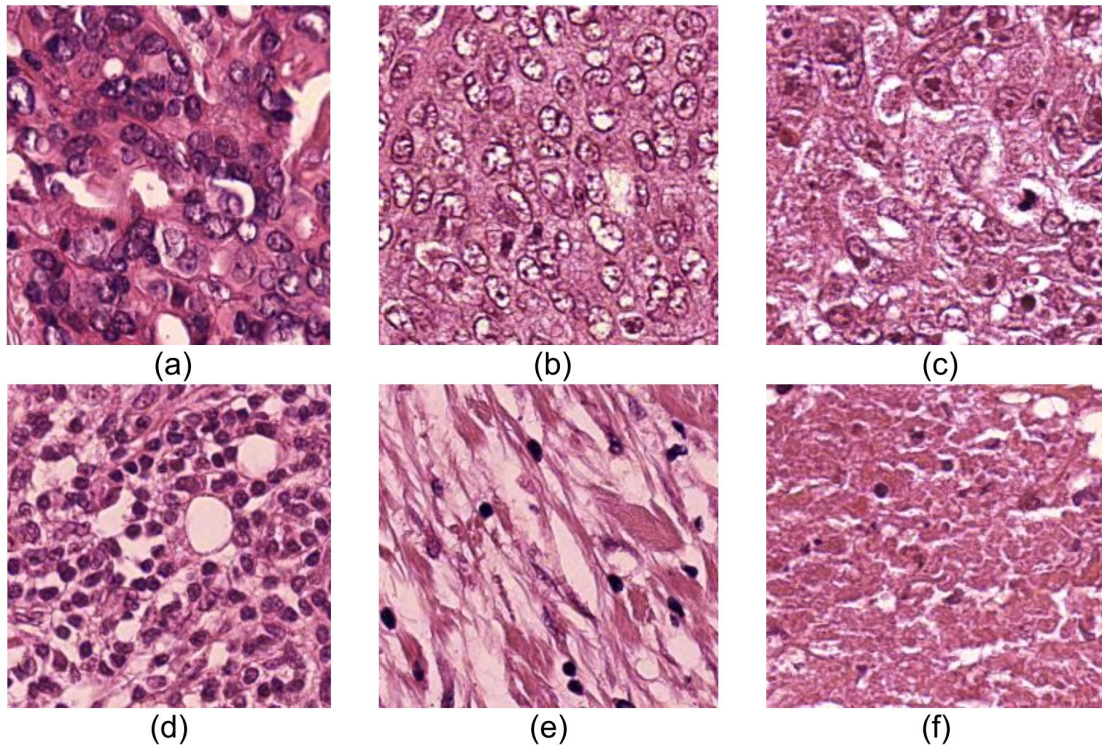


Figure 4.8: Samples of class-wise region crops from the slide images for the six-class classifier model designed. Scoring classes (SC) of nuclear atypia: (a) Score 1, (b) Score 2, (c) Score 3; Elimination classes (EC): (d) Lymphocytes, (e) Stroma, (f) Necrosis.

in their dimensions depending on the extent and distribution of the nuclei in slide images. Fixed-size image patches are extracted from these region crops based on detected nuclei seeds to train the CNN. Hence, the size variation of region crops is immaterial. Image patch size is an important parameter that has a significant impact on the performance of deep CNNs. Based on a set of experiments conducted with different patch sizes, it was found that a patch size of  $64 \times 64$  gives the best classification performance for the chosen CNN. The selection of patch size is explained further in Section 4.3.2.1. Class imbalance exists between slide images in the MITOS-ATYPIA dataset. The proportion of slide images that belong to score 2 class is very high compared to the other two classes. This results in class imbalance between extracted nuclear image patches as well. Among the scoring class patches, score 1 type patches are far less in number compared to score 2 type image patches. Hence, a set of basic rotation and flipping

Table 4.1: Performance comparison of different deep CNNs considered for the six-class classifier in the proposed framework.

<b>CNN Architecture</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>	<b>Accuracy</b>
MobileNetV2 (Sandler <i>et al.</i> , 2018)	0.8418	0.8403	0.8406	0.9517
Resnet50 (He <i>et al.</i> , 2016)	0.8645	0.8644	0.8644	0.9548
VGG16 (Simonyan and Zisserman, 2014)	0.8681	0.8641	0.8646	0.9547
Resnet101 (He <i>et al.</i> , 2016)	0.8728	0.8682	0.8692	0.9561
DenseNet121 (Huang <i>et al.</i> , 2017)	<b>0.8875</b>	<b>0.8857</b>	<b>0.8858</b>	<b>0.9619</b>

operations are applied on the extracted patches of lower proportion classes to equalize the training samples in all the classes.

#### *Selection and configuration of deep CNN*

The major functional engine in the proposed framework is a deep learning-based classifier that accurately classifies the nuclear image patches into one of the six classes. The CNN classifier used in the deep learning stage has a great impact on the performance of nuclear atypia scoring. The framework offers flexibility to apply any high-performing CNN classifier into the process pipeline to improve the overall nuclear atypia scoring. Experimented are conducted with a set of popular classifier CNNs such as VGG16 (Simonyan and Zisserman, 2014), Resnet (He *et al.*, 2016), DenseNet (Huang *et al.*, 2017), and MobileNetV2 (Sandler *et al.*, 2018). The candidate CNNs are separately trained using nuclear image patches of dimension  $64 \times 64$  extracted from region crops of all the six classes (ECs and SCs) and tested with unseen test data. Table 4.1 presents the performances of the various CNN models trained this way. It is observed that DenseNet (Huang *et al.*, 2017) gives the best performance for all the four measures precision, recall, F1 score, and accuracy.

DenseNet is a popular CNN architecture for image classification. It is a sequential

concatenation of multiple dense blocks separated by transition layers. Unlike the conventional CNNs in which the information flow (feature map) is sequential from layer to layer, DenseNet introduced additional information flow within the dense blocks. A dense block is made of multiple layers such that the output of every layer is connected to all subsequent layers in the same block in a feed-forward manner. This architecture facilitates feature reuse and reduction in trainable parameters. Apart from these, DenseNet has other advantages like low computational requirement, immunity to vanishing gradient, etc., compared to other recent architectures. Original DenseNet architecture is designed for an image dataset of 1000 classes. The DenseNet121 configuration is customized to suit the requirement of the proposed method. The input image size is configured to  $64 \times 64$  after experimenting with multiple input sizes. Output classes are fixed to six as there are six classes in the problem formulation. ReLU activation function (Nair and Hinton, 2010) is used in the intermediate layers and Softmax for the final classification layer. Categorical cross-entropy is the loss function used. For weight optimization, Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of  $3e-4$ . Other hyperparameters used in the CNN model are: dropout rate = 0.3, batch size = 32, EPOCHs = 125. The model parameters are finalized by a set of experiments carried out with a random split of the MITOS-ATYPIA dataset. Later, the performance and consistency of the model are evaluated by five-fold cross-validation.

### **4.2.3 Postprocessing and Atypia Scoring**

Once the trained CNN model is ready, slide-level prediction of nuclear atypia score is performed in a fully automated way using the test set of slide images. Algorithm 1 and Algorithm 2 are followed in this process. Every test slide image goes through the preprocessing stage described earlier to obtain the color-normalized form and the corresponding nuclei seed image. Further processing for slide level nuclear atypia scoring is illustrated in Figure 4.9. Using the nuclei seed points, patches of dimension  $64 \times 64$  are extracted irrespective of the classes, and fed to the trained CNN model. The model

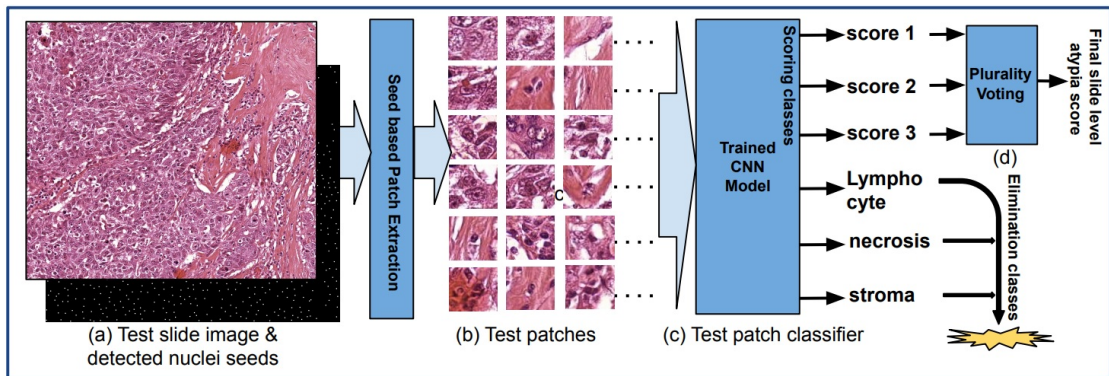


Figure 4.9: Slide level of evaluation of nuclear atypia score. Input to the evaluation pipeline is the preprocessed test instances consisting of color-normalized slide images and nuclei seeds detected. Output is the nuclear atypia score of the input slide image.

classifies every input nuclear image patch into one of the six classes. Samples classified into elimination classes (i.e., lymphocytes, necrosis, and stroma) are ignored as they are not involved in the scoring of nuclear atypia. A plurality voting scheme with priority-based tie-breaking (Algorithm 2) is applied to the scoring classes (score 1, 2, and 3) of nuclear image patches to assign the final nuclear atypia score for the slide image. In the case of equal votes obtained by two dominating score classes, the final score is assigned by giving preference to the higher score class. For example, if the number of nuclear patches in classes 'score 3' and 'score 2' are equal, and greater than 'score 1', the final score assigned to the parent slide image is 3. This is in line with the pathology procedure in practice. The elimination classes are defined for the CNN model to filter out all the non-malignant nuclei in the slide image that are irrelevant in the process of nuclear atypia scoring. In the absence of the elimination classes in the model, such nuclei present in the slide would have been classified into one of the scoring classes and adversely impact the performance of the model. In this way, the proposed method approximates a human

---

**Algorithm 1** Atypia score computation

---

```
1: procedure ATYPIA_SCORE_COMPUTE(SlideImage (I))
2:   Patch_size  $\leftarrow$  64
3:   I_seeds  $\leftarrow$  nucleiSeedDetection(I)
4:   I[n]  $\leftarrow$  imagePatchExtraction(I, I_seeds, Patch_size)
5:   [C1, C2, C3, ..., C6]  $\leftarrow$  patchClassPrediction(I[n])
6:   NAScoreI  $\leftarrow$  majorityVoting([C1, C2, C3])
7:   return NAScoreI
8: end procedure
```

---

---

**Algorithm 2** Plurality voting

---

```
1: procedure MAJORITY_VOTING(PatchCounts ([C1, C2, C3]))
2:   if (C3  $\geq$  C2 AND C3  $\geq$  C1) then
3:     NAScore  $\leftarrow$  3
4:   else if (C2  $>$  C3 AND C2  $\geq$  C1) then
5:     NAScore  $\leftarrow$  2
6:   else
7:     NAScore  $\leftarrow$  1
8:   end if
9:   return NAScore
10: end procedure
```

---

pathologist’s way of nuclear atypia scoring by looking only at the malignant tumor nuclei for atypia scoring. In Figure 4.10, samples of the voting pattern observed for different test images from both slide scanners are presented. The vote support for each class is given in brackets. In most of the cases, the voting majority of the winning class is large enough to clearly discriminate from the other two classes. In slide images, the total number of malignant cells/nuclei show variations as seen in these pie charts. A low total indicates that a major part of the image is dominated by non-malignant cells like lymphocytes, stroma cells, etc. A high number of malignant cells indicates the domination of tumor regions in the slide image.

### 4.3 Experimental Results & Discussions

The experiments on the proposed method have been carried out using the public dataset MITOS-ATYPIA. Most of the recent methods have used the same dataset for experi-

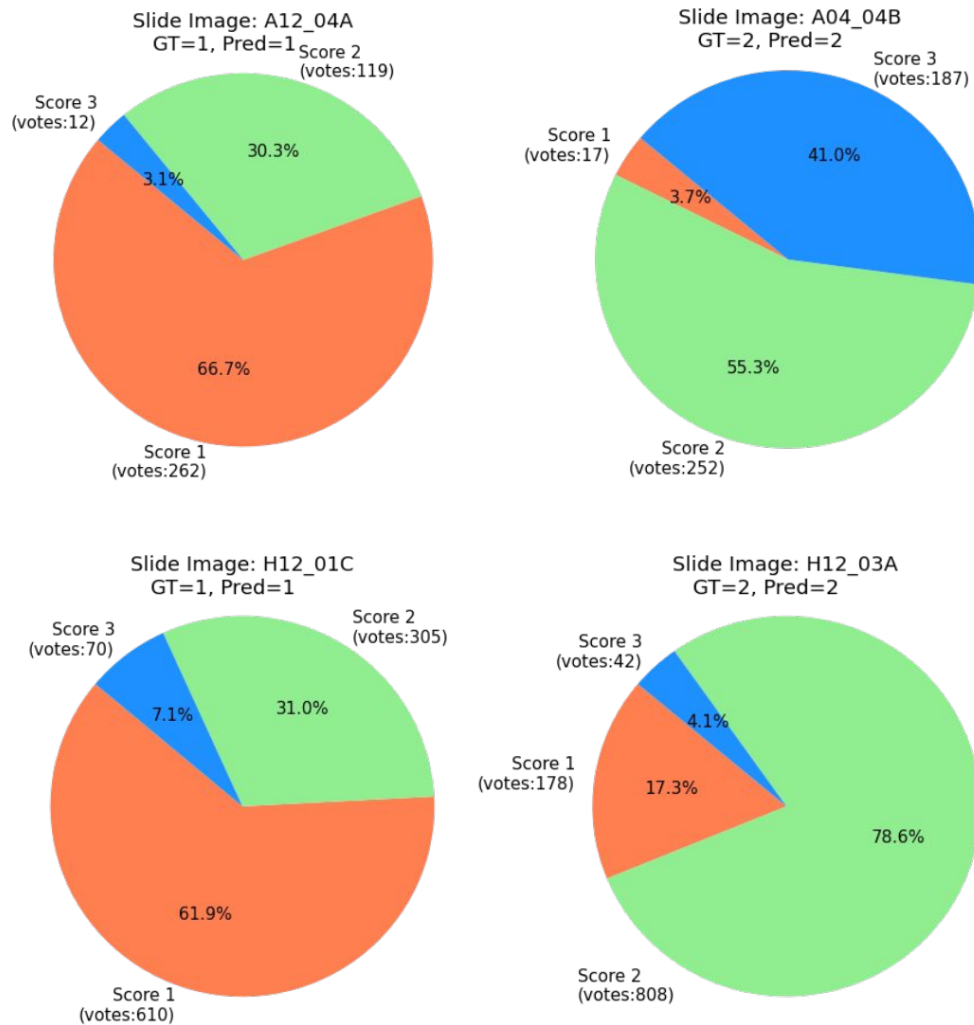


Figure 4.10: Instances of plurality voting-based prediction of nuclear atypia score from Aperio and Hamamatsu scanner image sets. (GT: Ground truth score).

ments and validation. This simplifies the performance comparison with these methods. In this section, first the details of the dataset, experimental setup, validation approach, and metrics are discussed. This is followed by the presentation of experiments, result comparison with existing methods, and a discussion on various aspects related to nuclear atypia scoring that are observed during this work.

### 4.3.1 Dataset & Experimental Setup

The MITOS-ATYPIA (ICPR, 2014) dataset is the only known large-scale public dataset exclusively for breast cancer atypia scoring. Most of the methods published after its release in the year 2014 have used this dataset. The dataset consists of 300 labeled slide images each from two different scanner models Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. The slide images are captured at a magnification of  $20\times$ . Every image in the Aperio scanner set has a dimension of  $1539 \times 1376$  pixels, whereas the Hamamatsu scanner images have a dimension of  $1663 \times 1485$  pixels. The ground truth label assigned to each slide image is an integer value of 1, 2 or 3 which corresponds to the nuclear atypia score of that image. The scores are assigned by experienced pathologists. Due to the commonly observed problem of pathologists' disagreement in nuclear atypia scoring, labeling of the MITOS-ATYPIA dataset was done in two stages. Initially, two pathologists independently assigned scores to every image in the dataset. If both gave the same score to an image, that is assigned as the final score for that image. In case of disagreement in the scores assigned for any image, a consensus score is assigned based on the opinion of a third pathologist. On analyzing the initial scoring by two pathologists it is found that out of the 22 subsets of images from both scanners together, eight subsets had more than 25% disagreement between the two pathologists. A senior pathologist who screened these subsets found that even the final scores assigned to images of the two subsets A11 and H11 are highly contentious and recommended excluding these subsets. Based on the recommendation, these subsets are excluded from the dataset. The dataset finally had 265 images each from both scanners, making a total of 530 labeled slide images.

#### *Nuclear image patch extraction and augmentation*

To design the CNN model parameters, a random split of the Aperio scanner image set is used with 80:20 ratio for training and testing. Image patch size  $64 \times 64$  is found to give the best classification accuracy for the six-class CNN. The MITOS-ATYPIA dataset has class imbalance among slide images of the three score classes, with score



1 and score 3 class images are proportionately less compared to score 2 class images. This is reflected in the extracted nuclei image patches as well. In a slide image there are approximately 200-300 nuclei present. On extracting the score 2 class nuclear patches from the Aperio scanner image set, the total number crossed 20,000 without any augmentation. Some samples are randomly removed from this set to round the number of score 2 class patches to 20,000. The number of such patches obtained for other classes are below 20,000. To equalize the number of training samples across all the classes, score 1, score 3, and the three elimination class image patches are augmented using rotation and flipping operations. That way 20,000 image patches are created per class to make a total of 1,20,000 training patches from the Aperio scanner image set. In the experiments conducted with slide images from both scanners together, nuclei image patches from both Aperio and Hamamatsu scanners are merged class-wise to create 40,000 samples per class. For the six classes together, a total of 2,40,000 nuclear image patches are used for training the CNN. The nuclei patch extraction and augmentation steps are automated using matlab scripts that take input as the slide image and nuclei seeds obtained from the nuclei seed detection stage of the proposed pipeline.

#### *Five-fold cross validation*

Since the MITOS-ATYPIA dataset contains images from two different scanners, most of the existing methods in literature have reported results based on three sets of experiments based on i) Aperio scanner images, ii) Hamamatsu scanner images, and iii) combination of both. For a fair comparison of the results, the same set of experiments are conducted for the proposed method also. Five-fold cross validation is performed in all the three sets of experiments to evaluate the consistency of the model. In each case, the corresponding dataset is split into five disjoint sets using Mersenne twister randomization algorithm (Matsumoto and Nishimura, 1998) to ensure that there is no bias in the splitting of the dataset. The five data folds are created from these splits, with one unique split forming the test set every time and remaining splits in the training set. This way every split becomes a test set exactly once to make sure that every slide image in

the dataset appears exactly once as a test image in any one of the five folds.

#### *Hardware and software setup*

The primary hardware used for training and evaluation of the proposed method is Tesla V100 GPU with 32 GB GPU memory. Both Matlab and Python libraries are used for the implementation of different stages. For deep learning, Keras Python library with Tensorflow as the backend is used.

#### *Evaluation metrics*

Class imbalance exists in the MITOS-ATYPIA dataset among the three scoring classes. The ratio of score 1, score 2, and score 3 class images in the dataset is 1:8:2. For training the CNN, data augmentation is applied to reduce the class-imbalance between the samples of different classes. But for the evaluation, class imbalance in the test set is retained to get the realistic performance of the proposed method. When there is a class imbalance, the appropriate metrics for evaluation of a classification system are precision, recall, and F1 score. These metrics are primarily used to measure the performance of the proposed method and for comparison. Apart from these, accuracy is also used for comparison with the existing methods. The details of these metrics are provided in Appendix A.

### **4.3.2 Results and Analysis**

Several analyses and experiments are conducted to design the model and evaluate the performance consistency of the proposed method. This subsection presents the outcome of the experimental analyses carried out and the observations. Initially, the performance of the six-class CNN model is presented briefly. This is followed by a detailed evaluation of the proposed method for slide image-level automated nuclear atypia scoring. The results are compared with state-of-the-art methods for automated nuclear atypia scoring.

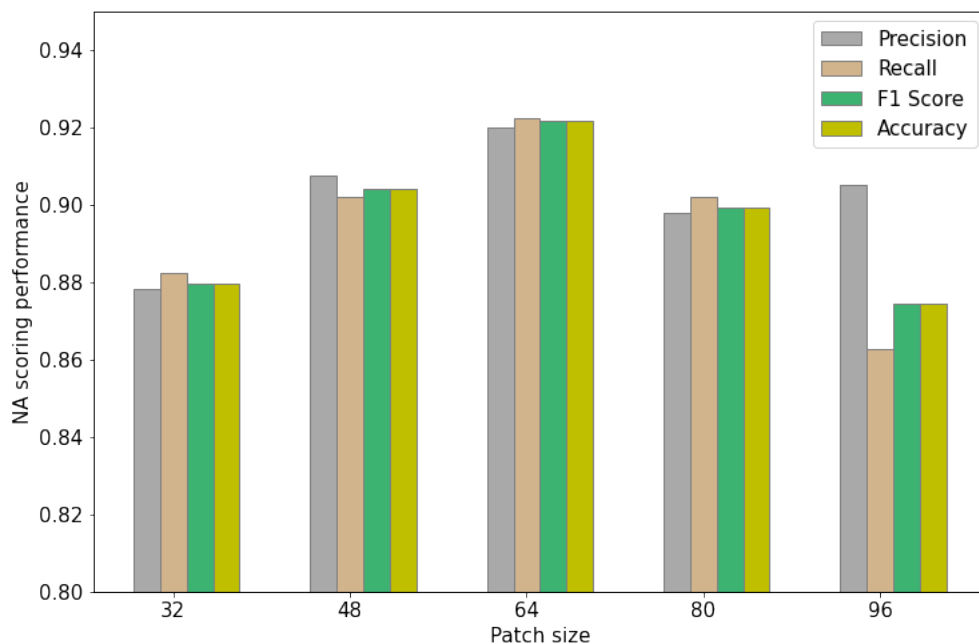


Figure 4.11: Result of experiments carried out to decide the input patch size to be used for the CNN.

#### 4.3.2.1 Selection of Patch Size for CNN

Image patch size has a significant impact on the performance of a CNN model based on patches extracted from a high-dimensional image. Appropriate patch size primarily depends on the task and nature of the dataset used. A practical approach to identify a suitable patch size for the model is to do it empirically. Experiments are conducted for this purpose on a random split of the dataset using different patch sizes such as 32, 48, 64, 80, and 96. DenseNet CNN is separately trained with nuclear image patches of these sizes and atypia scoring performance is evaluated. The results showed that a patch size of  $64 \times 64$  consistently gives the best performance on all the metrics considered. Figure 4.11 shows the performance pattern of nuclear atypia scoring with different patch sizes. On every metric, the performance shows an upward trend from patch size 32, and it peaks at 64, after which a diminishing trend is observed. On subjective analysis of the reason behind this pattern, it is observed that a desirable patch size has to correctly balance two aspects. One is that the patch size should be large enough to accommodate

Table 4.2: Performance of the nuclear image patch classifier model on different subsets of the MITOS-ATYPIA dataset.

Image subset	Precision	Recall	F1 Score	Accuracy
Aperio images (A)	0.8728	0.8695	0.8655	0.9565
Hamamatsu images (H)	0.8616	0.8584	0.8568	0.9528
Combined images (A & H)	0.8667	0.8701	0.8645	0.9547

even the largest nuclei (typically score 3 type nuclei) along with the cytoplasm and other immediate adjacent structures (context of the nuclei). On the other hand, if the patch size is too large, the patches extracted automatically during the testing process are likely to have more class overlapping (presence of different class nuclei in an image patch), causing more misclassifications. The empirically chosen patch size of 64 is observed to balance these two aspects to give the peak performance at this patch size.

#### 4.3.2.2 Performance of the CNN Classifier

Effectiveness of the DenseNet model in classifying the six classes of nuclear image patches is evaluated using five-fold cross-validation on Aperio, Hamamatsu and the combined image sets. Table 4.2 shows the nuclear patch-level classification performance of the six-class CNN model. For training the CNN, nuclear image patches of dimension  $64 \times 64$  are extracted from region crops of all the six classes. These nuclear image patches are extracted from manually annotated non-overlapping class-wise region

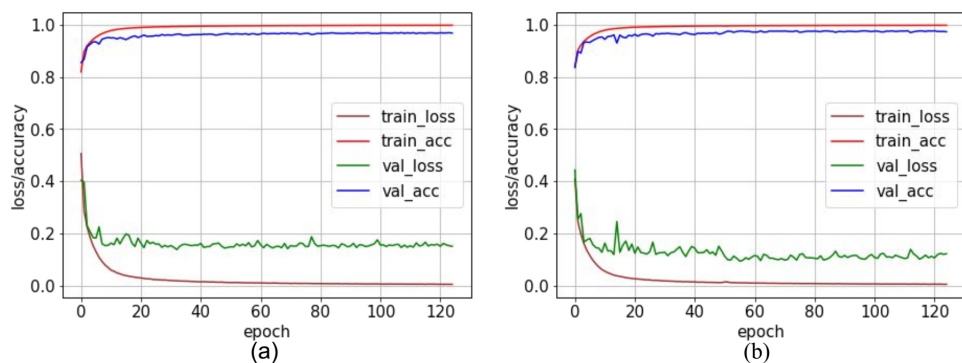


Figure 4.12: Learning pattern of the six-class CNN classifier (DenseNet) used in the proposed method for images sets from (a) Aperio scanner, (b) Hamamatsu scanner.

crops. The CNN classifier performance indicates how well the model can discriminate between nuclei belonging to different classes. The loss/accuracy graph in Figure 4.12 shows the learning pattern of the six-class CNN classifier for Aperio and Hamamatsu scanner image sets. The curves show a usual learning pattern with a few subtle differences between the two scanner image sets. Validation loss and accuracy pattern of Hamamatsu scanner images looks slightly better than that of the Aperio scanner image set. However, loss and accuracy show higher degree of fluctuations in the case of the model trained on Hamamatsu scanner images. On the contrary, the slide level evaluation of the model trained with Hamamatsu scanner images showed noticeably lower performance despite the higher validation accuracy observed in training. This indicates a slight over-fitting of the model to the Hamamatsu training set.

#### **4.3.2.3 Evaluation and Comparison of Nuclear Atypia Scoring**

The ultimate objective of the proposed framework and the method is to perform automated nuclear atypia scoring from histopathology slide images. Every step in the method pipeline such as image normalization, region crop extraction (for training), nuclear image patch classifier, etc., contributes to this final objective. This subsection presents how far the objective is achieved and how the proposed method fares in comparison with the state-of-the-art. To comply with the convention followed in most of the existing methods, three sets of experiments are conducted to evaluate the performance of slide image-level nuclear atypia scoring on i) Aperio scanner images, ii) Hamamatsu scanner images, iii) the combined set of images from both scanners. Five-fold cross-validation is performed in each of these cases, and average results are presented. The performance of the proposed method is compared with the recent methods reported in literature.

In the slide level evaluation, all nuclear image patches are extracted from the slide images in a fully automated way without any class consideration. So, these patches may have overlapping classes and more challenging scenarios than in the training phase. Per-

Table 4.3: Results of the proposed method and comparison with the state-of-the-art methods using Aperio scanner images.

Method	Precision	Recall	F1 Score	Accuracy
Xu <i>et al.</i> (2017)	0.7352	0.6646	0.6879	0.8000
Rezaeilouyeh <i>et al.</i> (2016)	0.2500	0.3333	0.2857	0.7500
Khan <i>et al.</i> (2015)	0.7851	0.7020	0.7197	0.8293
Lu <i>et al.</i> (2015)	0.4455	0.4551	0.4476	0.7800
Das <i>et al.</i> (2019)	0.8237	0.7196	0.7463	0.8533
Das <i>et al.</i> (2020a)	0.7623	0.7853	0.7901	0.8533
Das <i>et al.</i> (2018)	0.8328	0.8501	0.8409	0.9062
Proposed method	<b>0.8867</b>	<b>0.8860</b>	<b>0.8835</b>	<b>0.9261</b>

Table 4.4: Results of the proposed method and comparison with the state-of-the-art methods using Hamamatsu scanner images.

Method	Precision	Recall	F1 Score	Accuracy
Xu <i>et al.</i> (2017)	0.6899	0.6925	0.6838	0.7973
Rezaeilouyeh <i>et al.</i> (2016)	0.2492	0.3333	0.2852	0.7475
Khan <i>et al.</i> (2015)	0.7831	0.6507	0.6832	0.8189
Lu <i>et al.</i> (2015)	0.4234	0.3333	0.4023	0.7576
Das <i>et al.</i> (2019)	0.8225	0.6926	0.7196	0.8516
Das <i>et al.</i> (2020a)	0.7684	0.7971	0.7815	0.8649
Das <i>et al.</i> (2018)	0.7729	0.7889	0.7803	0.8784
Proposed method	<b>0.8568</b>	<b>0.8511</b>	<b>0.8504</b>	<b>0.9007</b>

formance comparison of slide image-level nuclear atypia scoring using Aperio scanner images is presented in Table 4.3. The metrics precision, recall, F1 score, and accuracy are used for the comparison. Sometimes precision and recall have a balancing effect between them. Hence F1 score, the harmonic mean of precision and recall, is considered as the reliable single metric indicator. For all the metrics the proposed method gives significant improvement over the state-of-the-art methods on Aperio scanner image set. The quality of the Aperio scanner images is found to be better with a balanced color expression. This has a positive impact on the results. In Table 4.4, the results of the proposed method on Hamamatsu scanner images are presented and compared with existing methods. In this case also the proposed method gives significant improvement over the existing methods on all the four metrics considered. Images from this scanner are characterized by their over-color expression (Figure 4.5(b)) and considered more

Table 4.5: Results of the proposed method and comparison with the state-of-the-art methods using combined Aperio and Hamamatsu scanner images.

<b>Method</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>	<b>Accuracy</b>
Xu <i>et al.</i> (2017)	0.6683	0.7649	0.7065	0.7987
Rezaeilouyeh <i>et al.</i> (2016)	0.2496	0.3333	0.2824	0.7487
Khan <i>et al.</i> (2015)	0.7676	0.6771	0.7085	0.8329
Lu <i>et al.</i> (2015)	0.4572	0.4340	0.4346	0.7705
Das <i>et al.</i> (2019)	0.8254	0.7656	0.7214	0.8530
Das <i>et al.</i> (2018)	0.7694	0.7971	0.7815	0.8658
Proposed method	<b>0.8766</b>	<b>0.8760</b>	<b>0.8745</b>	<b>0.9174</b>

challenging. There is a consistent performance decline for the Hamamatsu image set in all the listed methods compared to the corresponding result of the Aperio image set. However, the performance decline for the proposed method is around 3% whereas the existing best method (Das *et al.*, 2018) has nearly 6% decline. This indicates that the proposed method shows more resistance to color variations in the slide images from different sources or scanners.

Finally, Table 4.5 presents and compares the performance of the proposed method on the combined dataset from both scanners (i.e., the MITOS-ATYPIA dataset). Therefore, this can be considered as the final verdict on the performance of automated nuclear atypia scoring. As one would expect, the result values on the combined image set lie between the results of the individual scanner image sets (e.g., F1 score on Aperio set: 0.8835, Hamamatsu set: 0.8504, combined set: 0.8745). In the proposed method, the combined result is consistently above the average of individual results, pointing to the possible effect of an increase in the number of training samples in the combined dataset. The result values of the proposed method have exceeded the state-of-the-art methods by a significant margin. The F1 score of the proposed method has 11.90% improvement over the best method in the literature (Das *et al.*, 2018). The corresponding precision and recall values of the proposed method are also improved by 13.93% and 9.89% respectively.

The receiver operating characteristic (ROC) graph is a commonly adopted way to

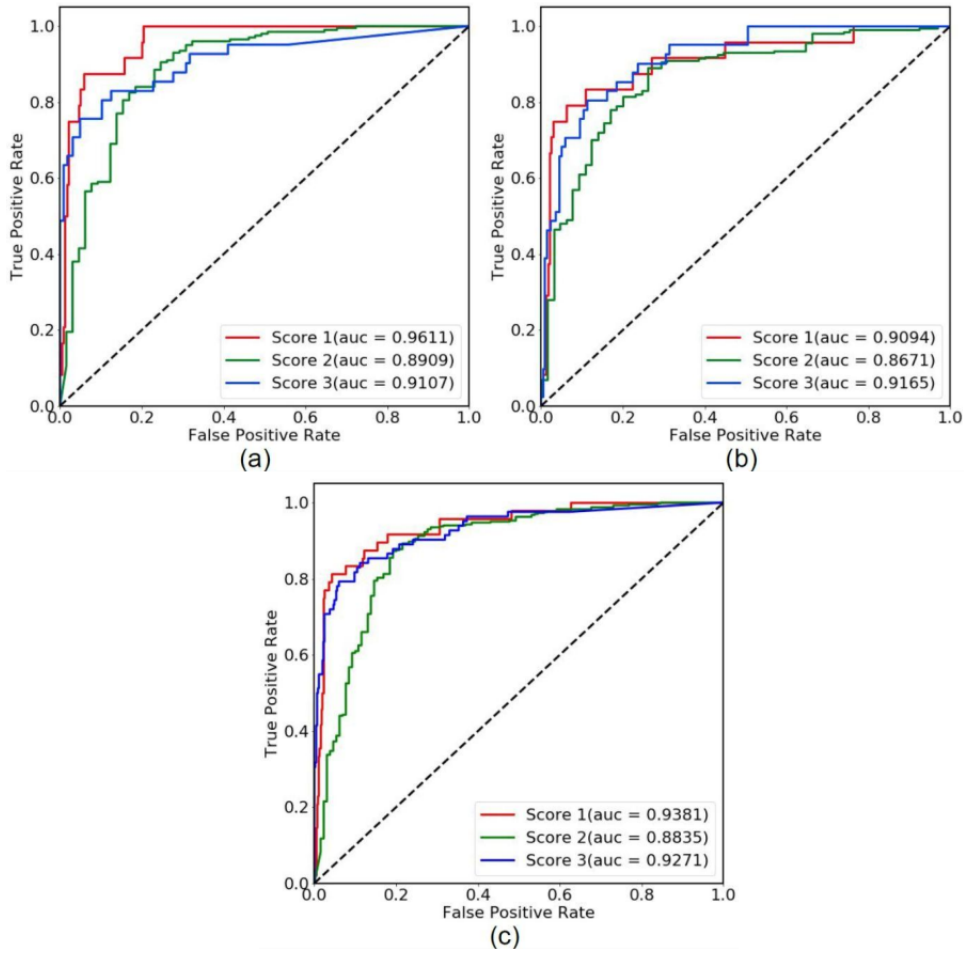


Figure 4.13: ROC curve and AuC for the three scoring classes of tumor cells for image set (a) Aperio, (b) Hamamatsu, (c) Combined MITOS-ATYPIA dataset.

show the effectiveness of classification systems. In the ROC graph, the true positive rate (TPR) is plotted against the false positive rate (FPR) at all classification thresholds to obtain the curves for each class. The area under the ROC curve (AUC) is an aggregate measure of classification performance at all possible thresholds. Figure 4.13 shows the ROC graphs for the proposed method with the three image sets. For each scanner image set, ROC curves of the three atypia classes are plotted and corresponding AUC is shown. The maximum value of AUC is 1.0 which represents a perfect classifier. AUC is threshold-invariant and scale-invariant that makes it a desired measure for classification systems. The final AUC is obtained by averaging class-wise AUCs and compared with the state-of-the-art methods in Table 4.6. Consistent with the improvements in



Table 4.6: Comparison of average area under ROC curve (AUC) with state-of-the-art for different scanner image sets.

Method	Aperio images	Hamamatsu images	Combined images
Xu <i>et al.</i> (2017)	0.5913	0.5986	0.5954
Rezaeilouyeh <i>et al.</i> (2016)	0.5922	0.5969	0.5945
Khan <i>et al.</i> (2015)	0.7564	0.7532	0.7554
Lu <i>et al.</i> (2015)	0.6123	0.6106	0.6114
Das <i>et al.</i> (2019)	0.8642	0.8639	0.8644
Proposed method	<b>0.9209</b>	<b>0.8977</b>	<b>0.9162</b>

precision, recall, F1 score, and accuracy presented before, AUC values obtained for the proposed method are also improved substantially over the existing methods.

### 4.3.3 Discussion

In this subsection, some of the analyses conducted during the development of the proposed method and the observations from those are discussed.

#### 4.3.3.1 Pathologists' Disagreement and Labeling Discrepancies

Inter-observer variability and reproducibility are two common issues in manual pathology procedures (Robbins *et al.*, 1995; Frierson Jr *et al.*, 1995; Dalton *et al.*, 2000). Even between pathologists who are experts in the domain, there can be disagreements on the measurements or scores assigned to specimens under analysis. Often this is resolved by plurality voting of a group of pathologists who perform independent evaluations. For the images in the MITOS-ATYPIA dataset, score labels are independently assigned by two senior pathologists. For the cases where there is no consensus between these two pathologists, a third pathologist's opinion is taken to assign a final score. Slide images in the dataset are grouped into 11 subsets. The criteria for this subset grouping are unspecified, but images in a subset share similar visual appearance as if they are collected from the same biopsy slide, and perhaps from the same patient as well. The degree of disagreement between the two senior pathologists in the initial score they assigned

Table 4.7: Degree of disagreement between the two pathologists in the first level scoring of the MITOS-ATYPIA Aperio scanner images. The subsets A10 and A14 show no disagreement whereas A18 shows scoring disagreement between the two pathologists for 50% of the images.

Data subset	A03	A04	A05	A07	A10	A11	A12	A14	A15	A17	A18
Pathologists' disagreement	4%	31%	7%	25%	0%	28%	14%	0%	13%	5%	50%

to each slide image in the dataset is analyzed with respect to these subsets. Table 4.7 shows the result of this analysis on the Aperio scanner image set.

Subset A18 had the highest disagreement of 50% between the two pathologists. The third pathologists labeled all images in this subset with the same final score of 2. Cohen's Kappa score (Cohen, 1960), a metric for inter-annotator agreement, for the first two pathologists is computed as 0.62 which is moderate (value 1.0 for perfect agreement). When the independent scoring by two senior pathologists shows such a high degree of disagreement, there is a strong possibility of error in the final score assignment as well. In our case, all the subsets with a disagreement  $\geq 25\%$  are screened by a senior pathologist. In this screening, the final score labels assigned to the subsets A04, A07, and A18 are found to be reasonable with minor discrepancies, whereas final score labels of images in the subset A11 are found to remain contentious. Figure 4.2(c) and Figure 4.2(d) shows the similarity between extracts from two slide images of subset A11 labeled with score 2 and score 3 respectively. These two tumor regions are hardly differentiable even for an experienced pathologist. On the recommendation of this senior pathologist, A11 and the corresponding subset H11 of Hamamatsu scanner are removed from the dataset. The remaining 20 subsets are used in the proposed method for experiments. The observation is that the performance of automated methods can be negatively impacted if the models are trained with datasets having labeling discrepancies. This will be counterproductive for the research efforts put on the related tasks.

Table 4.8: Comparison of the proposed method with the initial scoring of the MITOS-DATASET by two independent pathologists.

<b>Evaluator</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>	<b>Accuracy</b>
Pathologist 1	0.9452	0.9360	0.9379	0.9574
Pathologist 2	0.8973	0.8990	0.8980	0.9078
Proposed method (Aperio images)	0.8867	0.8860	0.8835	0.9261
Proposed method (Hamamatsu images)	0.8568	0.8511	0.8504	0.9007
Proposed method (Combined set)	0.8766	0.8760	0.8745	0.9174

#### 4.3.3.2 Nuclear Atypia Scoring: Man vs. Machine

Comparing the performance of automated nuclear atypia scoring with the scoring by human pathologists can give some interesting insights. Considering the practical difficulties of getting all slide images freshly assessed by a pathologist, the initial scoring of the MITOS-ATYPIA dataset by two senior pathologists is taken as samples for comparison with the proposed method. Reference for this comparison is the final score assigned to the images with the help of the third pathologist. Table 4.8 shows the result of this comparison. Apparently, the results of the proposed method have not surpassed the human pathologists, but they are close. Comparing the nearest results, the difference between scoring by pathologist 2 and the proposed method on Aperio scanner images is in the range 1%-2%. If the possible image annotation discrepancies are eliminated, the proposed method stands a good chance to beat human pathologists. Apart from that, by improving the algorithms in different phases of the framework, i.e., by using better normalization, nuclei detection, and deep learning algorithms, the results of the proposed method can be further improved. With the current performance, the proposed method is quite useful as an assistive tool to reduce the workload of pathologists, by using prediction probability to selectively review the method's prediction and make corrections if required.

One of the notable features of the proposed method is that it nearly emulates the

routine manual procedure followed by a pathologist for nuclear atypia scoring with help of the vision capability of a deep learning algorithm. In manual atypia scoring, pathologists focus on the malignant cells in the tumor region and ignore the cells like lymphocytes, stroma cells, etc. By classifying nuclei as scoring classes and elimination classes, the method considers only malignant tumor cells for atypia scoring and ignores the rest. In clinical practice, a pathologist does a subjective assessment of the size, shape, and extent of tumor nuclei to assign the atypia score. In the automated method, size and shape variations of different nuclei classes are learned by the deep learning algorithm from nuclear image patches to classify the nuclei. The plurality voting followed in the proposed method approximates the assessment of the extent of each type of tumor nuclei by a pathologist. Overall, an effective approximation of the routine pathology procedure can be considered as the major factor behind the excellent performance improvement in the proposed method.

#### **4.3.3.3 Future Prospects for the Proposed Framework**

The proposed framework for nuclear atypia scoring offers flexibility in each of its major phases namely preprocessing, deep learning, and postprocessing. The preprocessing phase is aimed at color normalization and nuclei detection in H & E histopathology images. There are several existing methods for these operations. Moreover, research is still going on to develop better algorithms. The proposed method used two well-known methods (Reinhard *et al.*, 2001; Al-Kofahi *et al.*, 2009) in this phase. These methods can be replaced with better performing algorithms to create a method pipeline that can further improve the atypia scoring performance. As mentioned before, the backbone of the proposed framework is the deep CNN which accurately classifies the nuclear image patches. Deep learning is an aggressively researched area and new algorithms are quite frequently proposed in literature. Deep learning phase of the proposed framework is envisaged to fit any CNN classifier as the algorithm in this phase. The method uses DenseNet (Huang *et al.*, 2017) for this purpose after comparing the performance with

several existing CNNs (Table 4.1). In future, if this CNN is replaced with a better one, the performance of nuclear atypia scoring is bound to improve from what is achieved by the proposed method. The postprocessing phase is for the prediction of slide level atypia score using the output of CNN classifier. CNN outputs the distribution of nuclear image patches into the six classes. A plurality voting scheme with priority based tie-breaking is used (Algorithm 2) on the scoring classes to predict the final atypia score. It is possible to replace this algorithm with any other scheme or algorithm that does the final atypia scoring more accurately. In a nutshell, the formulation of slide level nuclear atypia scoring as a nuclei classification problem in this framework and effective utilization of suitable algorithms at different phases resulted in the excellent performance improvement in nuclear atypia scoring. The flexibility in the framework makes it promising for further development of automated nuclear atypia scoring.

## **4.4 Summary**

Manual nuclear atypia scoring is tedious, error-prone, and has low reproducibility due to its subjective nature. Automating this pathology procedure through image analysis has been attempted by many in recent years. It is a challenging task to mathematically model the histopathology images due to their structural complexity and diversity. The attempts towards this in the existing methods have not resulted in great performance or generalizable solutions. In this chapter, a novel deep learning-based framework for automated nuclear atypia scoring is presented. This framework offers the flexibility to apply different algorithms in its various phases to create new method pipelines with potential for performance improvement. The intrinsic three-class problem of slide level nuclear atypia scoring is reformulated as a six-class problem of nuclei classification in the deep learning phase of the framework. This formulation aids the effective use of deep CNNs to classify all the nuclei present in slide images. This classification is utilized in the post-processing phase for accurate prediction of atypia score. The proposed

method follows this framework and gives results that are significantly above the state-of-the-art. The flexible nature of the framework in its deep learning phase is highly relevant considering the rapid advancements happening in deep learning. Any existing or emerging deep CNN with better classification performance can be easily configured in the framework to improve the results. The proposed framework and the strategy adopted in it are capable of taking automated nuclear atypia scoring closer to application in routine clinical practice.

## CHAPTER 5

# AUTOMATED MOLECULAR SUBTYPING OF BREAST CANCER

In this chapter, a deep learning-based classifier framework for automated molecular subtyping of breast cancer is proposed. IHC slide images of the four biomarkers are separately processed by the proposed framework. In the preprocessing stage, the non-informative background regions from the images are separated. The patches extracted from the foreground regions are classified into target classes using convolutional neural networks (CNN) models trained for this purpose. Classification results are post-processed to predict the status of all the four biomarkers. The predictions for the individual biomarkers are finally consolidated as per clinical guidelines to determine the subtype of the cancer. The proposed system is evaluated for the performance of biomarker status predictions and patient-level subtype classification. In both these aspects, the results obtained are promising. The F1 score values obtained for ER, PR, HER2, and Ki67 status assessment are 1.00, 1.00, 0.90, and 0.86 respectively. For patient-level molecular subtype classification, our method obtained an F1 score of 0.89.

### 5.1 Introduction

Breast cancer has several classifications (Malhotra *et al.*, 2010) based on different factors such as histological, molecular, functional etc. Invasive ductal carcinoma (IDC), a histological subtype of breast cancer constitutes 80% of all the breast cancer cases (Weigelt *et al.*, 2010). A therapeutically relevant classification of breast cancer is the molecular subtyping. Molecular subtype of breast cancer is determined based on the expression of protein biomarkers namely estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and antigen Ki67 in tumor tissues. The common molecular subtypes of breast cancer are Luminal A, Luminal B, HER2-enriched, and Triple-negative/Basal-like (Prat *et al.*, 2015; Tsang and Tse, 2020).

Table 5.1: Molecular subtypes of breast cancer and their characteristics (Table created from Eliyatkin *et al.* (2015))

<b>Molecular Subtype</b>	<b>Biomarker response</b>	<b>Prognosis</b>	<b>Targeted therapies</b>
Luminal A	ER+, PR+/-, HER2-, low Ki67	Good	Hormone therapy
Luminal B	ER+, PR+/-, HER2+, high Ki67	Intermediate	Hormone therapy
HER2 Enriched	ER-, PR-, HER2+	Worse	HER2 targeted therapies
Triple -ve	ER-, PR-, HER2-	Worse	Under investigation

The biomarkers ER and PR are hormonal receptors that promote cell growth and replication in the presence of the hormones estrogen and progesterone respectively. Hence, tumor cells with ER or PR grow and multiply aggressively in the presence of these hormones. HER2 is another protein receptor found in the cell surface and facilitates cell growth. Some tumor cells produce HER2 in large amounts and cause accelerated growth of tumors. Over-production of one or more of these protein receptors are triggered by the underlying genetic mutations that lead to abnormal cell proliferation and tumor formation. The antigen Ki67 is normally present in cells that are in the different stages of division. Hence, it is possible to estimate the cell proliferation rate by the IHC analysis of Ki67. Combining the responses of these four biomarkers (ER, PR, HER2, and Ki67), the molecular subtype of breast cancer is determined. Table 5.1 shows four major molecular subtypes of breast cancer and the characteristics of these subtypes with respect to biomarker response, prognosis, and treatment approach. Molecular subtyping of breast cancer requires assessment of all the four biomarkers for a patient.

Automated molecular subtyping of breast cancer involves processing of IHC biomarker images of ER, PR, HER2, and Ki67 (Figure 1.2) from the same patient. The extensive search in the existing literature shows that there are no such automated methods available currently for immunohistochemistry based molecular subtyping of breast cancer. A dataset that contains IHC images of ER, PR, HER2, and Ki67 collected patient-wise is not known to exist in the public domain. This can be a reason for the non-existence of prior automated methods for molecular subtyping of breast cancer. In



this work, the attempt is to automate this procedure using a patient-level dataset prepared by a collaborating medical research institute.

Deep learning algorithms are proven to be effective in several medical image analysis tasks including pathology image analysis (Litjens *et al.*, 2017; Niazi *et al.*, 2019). When there are sufficient labeled samples to train the models, deep learning algorithms like CNNs are quite effective in learning the inherent features from the training samples. The models trained in this way can classify unseen samples accurately. In this chapter, a novel classifier framework is proposed to analyze IHC images of all the four biomarkers collected from each patient to predict the molecular subtype of breast cancer. Status of each biomarker is predicted using a separate CNN based process pipeline. Due to the high dimension of IHC images, a patch-wise approach is adopted to train the CNN models. The process pipelines predict the status of the biomarkers as ER +ve/-ve, PR +ve/-ve, Ki67 low/high, and HER2 +ve/-ve/equivocal. Finally, the status of the four biomarkers are combined to predict the molecular subtype of breast cancer. Accurate determination of molecular subtype helps oncologists to decide the targeted treatment plan for a patient.

The major contributions of this work are as follows:

- A novel deep learning-based classifier framework is proposed for molecular subtyping of breast cancer using immunohistochemistry image analysis. Currently there is no such method found in literature.
- Our method resembles the manual pathology procedure for molecular subtyping in a manner that the biomarker statuses are computed separately for all the four biomarkers involved (ER, PR, HER2, & Ki67) and then combined as per clinical guidelines to determine molecular subtype.
- The proposed method provides high performance in slide level and patient-level prediction of biomarker status for all four biomarkers and that results in accurate molecular subtyping. This automated approach has the potential to reduce the workload, time-delay, and cost associated with the manual pathology procedure for molecular subtyping.

The remaining sections of this chapter are organized as follows. The detailed explanation of the deep learning-based framework proposed for molecular subtyping is

presented in Section 5.2. The experimental analysis carried out to validate the proposed method is presented in Section 5.3. Finally, the chapter is concluded with a discussion on the future prospects and clinical application of automated molecular subtyping.

## **5.2 Methodology**

Automated molecular subtyping of breast cancer requires the determination of all the four biomarker responses from digitized IHC slide images. For this purpose, the slide images of each biomarker are processed separately. In the proposed method, assessment of each biomarker's response is modeled as a classification problem. The target classes are defined by the biomarker responses that are essentially required for molecular subtyping. The final status of hormone receptors ER and PR are either positive or negative (ER+/- and PR +/-). For Ki67, the assessment is done to check whether Ki67 presence is low or high based on the extent of nuclei showing positive Ki67 response. HER2 response evaluation results in three outcomes namely Positive, Negative, and Equivocal.

The methodology has two major parts namely i) Training pipeline, ii) Evaluation pipeline. In the training pipeline, four separate CNN classifier models are trained to classify the image patches extracted from IHC images of ER, PR, HER2, and Ki67. Evaluation pipeline processes the unseen IHC images of these biomarkers from a breast cancer patient to predict the status of each biomarker and the molecular subtype of the cancer in a fully automated way.

### **5.2.1 Training Pipeline**

The training pipeline uses the training set of IHC images to train four independent CNN models i.e., one for each biomarker. The stages involved in the training pipeline of the proposed method is shown in Figure 5.1. It has two major stages namely, a) Image patch

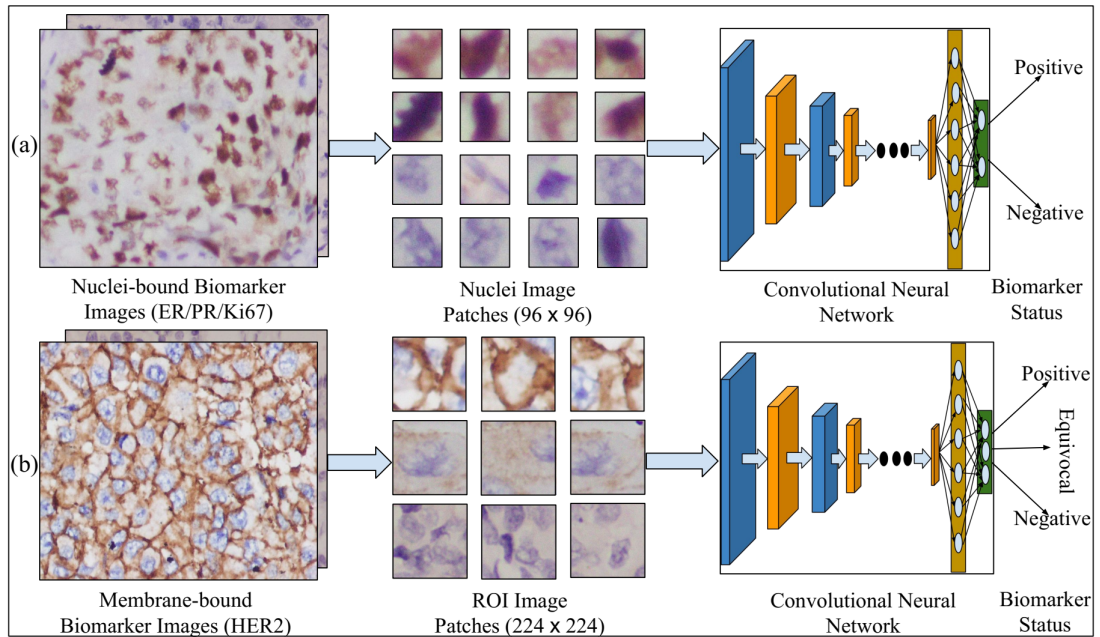


Figure 5.1: Training pipeline of the proposed framework for molecular subtyping. a) ER, PR, & Ki67 images are used to train three binary CNN classifier models separately for each biomarker. b) For HER2, a three-class CNN model is trained to classify each region patch into one of the output classes.

extraction b) Training the CNN.

### 5.2.1.1 Image Patch Extraction & Augmentation

Normally, IHC slide images have large spatial dimensions with plenty of background or non-informative regions present. Feeding such large images to CNN-like algorithms is computationally expensive. Hence, for training the deep learning models an approach of cropping small size image patches containing relevant information is adopted. In this stage, the image patches of fixed dimension are extracted from the slide images. Due to the inherent nature of the biomarkers, the process pipeline for ER, PR, & Ki67 are slightly different from the one used for HER2 images.

In the case of ER, PR, & Ki67, the problem is formulated as a binary classification of nuclear image patches. For this, nuclei-centric patches of dimension  $96 \times 96$  of both positive and negative classes are extracted from the training images. Class-wise label-

ing of the nuclei performed under the supervision of a senior pathologist is used for the extraction. Varying size of the nuclei is a challenge in deciding an appropriate patch size. In this case, the patch size is determined by considering the average pixel area per nucleus in the IHC images such that a sufficiently large portion of one nucleus or the complete nucleus is included in all the patches extracted. Simple rotations of  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$  are applied on the extracted patches to equalize the negative and positive class samples used for training. Being a membrane-bound receptor, color response for HER2 is visible in the cell membrane. The color intensity and the completeness of the brownish boundary formed around nuclei are considered by pathologists in manual evaluation of HER2 response. Taking this into consideration, larger image patches of dimension  $224 \times 224$  are extracted using the class-wise rectangular region labeling provided for HER2 images. HER2 response is formulated as a three-class problem where each patch is classified as positive, negative, or equivocal. Since the slide images of the equivocal class are less, a sliding-window based extraction of the patches along with the rotation operations is applied to create sufficient augmented samples for training.

### 5.2.1.2 Training of CNN Models

The proposed automated molecular subtyping framework has used convolution neural network (CNN), a powerful class of deep learning algorithms as the backbone. CNNs have been quite effective in several medical image analysis tasks including pathology image analysis (Litjens *et al.*, 2017). The popular CNN architecture namely DenseNet (Huang *et al.*, 2017) is customized to use in the proposed framework. DenseNet is a powerful classifier with less learnable parameters and a competitive training time. It is resilient to the vanishing gradient problem and supports feature reuse. These advantages and the prior experience with DenseNet are the motivation to choose this as the deep learning algorithm for classification of IHC image patches. However, the framework allows flexibility to replace the CNN module with any other classifier CNNs.

The original version of DenseNet is tested on benchmark datasets namely CIFAR-

Table 5.2: Parameters used for DenseNet CNN

<b>Configuration</b>	<b>Value/Function</b>
Loss function	Categorical cross entropy
Learning rate	$3e^{-4}$
Optimizer	Adam
Training Batch size	32
Dropout rate	0.3
No. of training EPOCHs	100

10, CIFAR-100, and ImageNet that contain 10, 100, and 1000 classes respectively. Its base variant DenseNet121 is customized as a binary classifier for ER, PR, and Ki67. Since HER2 has three target classes, the CNN is configured as a three-class classifier for image patch classification. The other custom configurations applied for DenseNet121 are shown in Table 5.2. The output of the training pipeline consists of four CNN models for ER, PR, Ki67, and HER2 image patch classification. These models are employed in the evaluation pipeline of the framework to achieve automated molecular subtyping.

## 5.2.2 Evaluation Pipeline

The evaluation pipeline of the proposed method is a fully-automated end-to-end workflow that takes all the four types of IHC images from the same patient and predicts the molecular subtype of breast cancer under investigation. Multiple stages involved in this pipeline are preprocessing, image patch classification, post-processing, and the molecular subtyping. Figure 5.2 shows the evaluation pipeline of the proposed method. The various stages in this pipeline are explained in the following sections.

### 5.2.2.1 Pre-processing of IHC Images

IHC slide images normally contain plenty of background regions that are irrelevant in pathology analysis. Processing such regions may consume the computational resources unnecessarily and affect the performance. In the preprocessing stage, such background

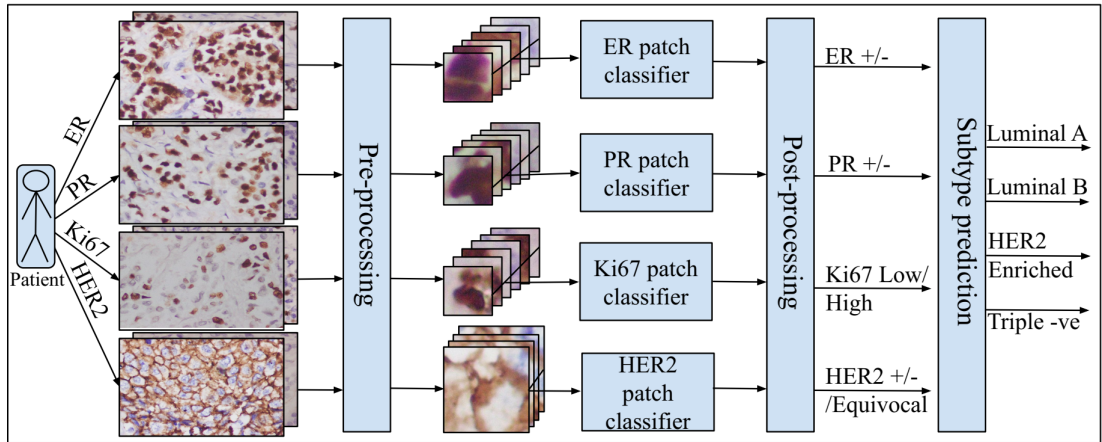


Figure 5.2: Evaluation pipeline of the proposed framework for automated molecular subtyping.

regions in the slide images are detected and excluded to focus on the objects/regions of interest that are primarily nuclei and cell membranes. The multi-otsu thresholding algorithm by Liao *et al.* (2001) is used to detect the background regions from IHC images of all four biomarkers. A binary image is obtained as an output to separate the foreground and the background. Further, morphological opening and closing operations are used to remove tiny non-cellular elements from the foreground. Figure 5.3 shows sample IHC images of the four biomarkers and the corresponding binary images showing foreground (in white) and background (in black). These binary images are used to extract only the nuclei and membrane region patches from IHC images for further processing. Subsequently, all foreground image patches of size  $96 \times 96$  are extracted from the ER, PR, and Ki67 images. From the membrane-bound HER2 images, patches of size  $224 \times 224$  are exacted as done in the training pipeline. Data augmentation is not done for these test samples to retain the class imbalance between samples in the evaluation process.

### 5.2.2.2 Image-patch Classification

The next stage is the classification of the patches from the biomarker images using the corresponding CNN models created in the training pipeline. In the case of ER, PR,

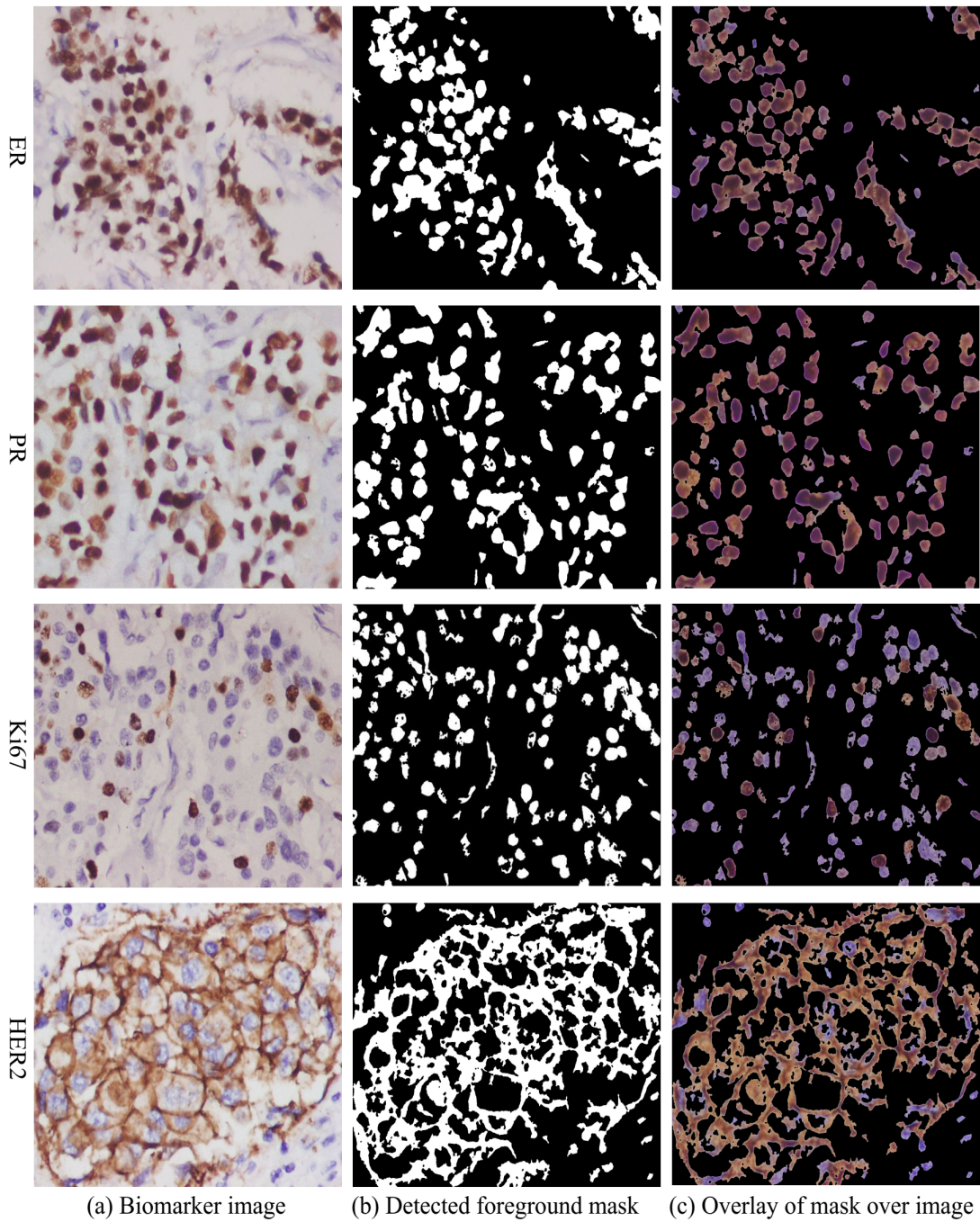


Figure 5.3: Extraction of foreground regions from the biomarker images to facilitate image patch extraction based on objects/regions of interest only: (a) Original IHC images of ER, PR, Ki67 & HER2; (b) Masks of detected foreground (white)/background (black); (c) Overlay of masks over the images.

and Ki67, each image patch is classified into a positive or negative class that indicates the biomarker response of the nucleus region included in the patch. HER2 patches are classified as positive, negative, or equivocal response classes using the three-class CNN model trained for this purpose. For each biomarker, the classification outputs are captured at slide level and patient level for further processing. The result of patch-level classification is fed to the post-processing stage for biomarker status predictions.

### 5.2.2.3 Post-processing for Biomarker Status Prediction

In this stage, the results of image patch classification obtained from the previous stage are processed to make prediction of biomarker status for all four biomarkers. The predictions are made at slide image level and patient level. In the slide image level prediction, the biomarker status of every slide image is predicted whereas in patient level prediction, the 10 slide images available for each biomarker per patient is considered to predict the biomarker status of ER, PR, Ki67, and HER2 for a patient.

---

#### Algorithm 3 Slide level assessment of biomarker status

---

```

1: procedure IHC_BIOMARKER_STATUS(SlideImage (Im), Biomarker model (Bm))
2:   Output: Biomarker response status
3:   Bg ← BackgroundDetector(Im)
4:   Impatch ← ForegroundPatchExtractor (Im, Bg, Bm.PatchSize)
5:   Pred ← Bm.CnnPrediction(Impatch)
6:   if (Bm.name = Er or Pr) then
7:     Bm.status = ErPrStatusCompute(Pred)
8:   else if (Bm.name = HER2) then
9:     Bm.status = Her2StatusCompute(Pred)
10:  else
11:    Bm.status = Ki67StatusCompute(Pred)
12:  end if
13:  return Bm.status
14: end procedure

```

---



---

**Algorithm 4** Determination of ER/PR status

---

```
1: procedure ERPRSTATUSCOMPUTE(Pred)
2:    $Frac = Pred[‘+ve’]/(Pred[‘+ve’] + Pred[‘-ve’])$ 
3:   if ( $Frac \geq 0.30$ ) then
4:     return ‘+ve’
5:   else
6:     return ‘-ve’
7:   end if
8: end procedure
```

---

---

**Algorithm 5** Determination of HER2 status

---

```
1: procedure HER2STATUSCOMPUTE(Pred)
2:   if ( $Pred[‘+ve’].Count \geq Pred[‘Eqv’].Count$ ) and
3:     ( $Pred[‘+ve’].Count \geq Pred[‘-ve’].Count$ ) then
4:     return ‘+ve’
5:   else if ( $Pred[‘Eqv’] \geq Pred[‘-ve’]$ ) then
6:     return ‘Eqv’
7:   else
8:     return ‘-ve’
9:   end if
10: end procedure
```

---

---

**Algorithm 6** Determination of Ki67 status

---

```
1: procedure Ki67STATUSCOMPUTE(Pred)
2:    $Frac = Pred[‘+ve’]/(Pred[‘+ve’] + Pred[‘-ve’])$ 
3:   if ( $Frac \geq 0.14$  and  $Pred[‘+ve’].Count > 20$ ) then
4:     return ‘High’
5:   else
6:     return ‘Low’
7:   end if
8: end procedure
```

---

The major steps involved in slide level biomarker evaluation is presented in Algorithm 3. In the case of ER and PR, a proportion threshold value is used to assign biomarker status (Algorithm 4). If the proportion of the positive patches classified by the respective CNN models is greater than or equal to 0.30, the biomarker status is assigned as positive (ER +ve/PR +ve). Otherwise, it is assigned as negative (ER -ve/PR -ve). The threshold value of 0.30 is derived from the Allred scoring system (Allred *et al.*, 1998; Harvey *et al.*, 1999) used for ER and PR. For HER2, a plurality voting strategy is applied (Algorithm 5) for the patches that are classified into the three classes

namely positive, negative, and equivocal. The Ki67 status used for molecular subtyping involves two subjective levels of Ki67 as low or high. A commonly used clinicopathology criteria in manual procedure is a threshold level of 14% positive Ki67 cells for assigning the Ki67 status. In the automated approach, the number of nuclei in each class are not counted, rather the image patches extracted based on the nuclei region obtained after background detection are used. The proposed method combined the 14% criteria with a threshold value for the number of positive class patches for Ki67 status prediction (Algorithm 6). Empirically it is found that the optimal threshold count value that gives the best prediction results is 20. This is further discussed in the experimental results section (Section 5.3.2.2).

#### **5.2.2.4 Molecular Subtyping**

The final stage in the proposed classification framework is the determination of the molecular subtype of breast cancer for a patient under evaluation. The input to this stage is the individual status of all four biomarkers involved i.e., ER, PR, Ki67, and HER2. Once the patient's level biomarker statuses are obtained from the previous stage, the molecular subtype of the cancer is determined in a rule-based manner using the criteria specified in Table 5.1. The cancer type is classified into one of the classes from Luminal A, Luminal B, HER2 Enriched, or Triple Negative. These are the most commonly used molecular subtypes of breast cancer.

### **5.3 Experimental Results & Discussion**

In this section, initially the dataset and experimental setup used for the development and validation of the proposed method are described. Subsequently, the experimental analysis carried out to validate the proposed method and the results are discussed.

### 5.3.1 Dataset and Experimental Setup

The major challenge in developing a fully automated method for molecular subtyping is the availability of a suitable dataset. At present, there are no public datasets which contain IHC images of all the four biomarkers from the same patients. Such a dataset is an essential requirement since molecular subtyping is a patient-level procedure. The IHC image dataset used for development of the proposed method is collected from the Department of Pathology, Kasturba Medical College, Mangalore, India. It consists of 800 IHC images of dimension  $1920 \times 1440$  collected from 20 breast cancer patients. All the images are captured at  $40\times$  magnification of the microscope. For each patient, there are 10 IHC images per biomarker that makes a total of 40 images per patient considering the four biomarkers (ER, PR, Ki67, and HER2) needed for molecular subtyping. Molecular subtypes of all the 20 patients, determined manually by the pathologists, are known and are used as the ground-truths for automated evaluation. In addition, slide-level and patient-level status of all four biomarkers are determined by a set of experienced pathologists and these are used to validate the proposed automated method in each level.

For the validation of the proposed method, a four-fold cross validation is used at each level of evaluation. The dataset of 20 patients is divided into four splits of five patients each. Four cross validation folds are created from these splits by taking one split at a time as test set and the remaining three splits together as the training set. Since there are 10 slide images per biomarker for every patient, the test set of every fold contains 50 slide images per biomarker and the training set contains 150 slide images per biomarker. The use of cross validation ensured that every patient sample appears in the test set once in any one of the four folds. This way the attempt is to make sure that the obtained results are not influenced by any bias in the selection of the test set. Image patches are extracted from the slide images to train the CNN. The details of training patches used for each biomarker is shown in Table 5.3. Basic augmentation techniques like translation and rotation are applied to equalize the number of samples in each class

Table 5.3: Details of the training image patches used to train CNN models used in the proposed method

<b>Biomarker</b>	<b>Patch Size</b>	<b>Patches Per Classes</b>
ER	96 × 96	Positive: 40,000, Negative: 40,000
PR	96 × 96	Positive: 40,000, Negative: 40,000
Ki67	96 × 96	Positive: 40,000, Negative: 40,000
HER2	224 × 224	Positive: 13,500, Negative: 13,500, Equivocal: 13500

of a biomarker.

The hardware configurations used for training and testing of the proposed method are IBM Power9 CPU, 2 Tesla V100 GPUs with 32GB GPU-RAM for each. The software frameworks used for the implementation primarily include Keras and Tensorflow.

### 5.3.1.1 Evaluation Metrics

There exists a class-imbalance in the dataset starting from the patient level with respect to the samples of different molecular subtypes. For example, the number of patient samples that belong to the molecular subtype Luminal A is seven whereas Luminal B has only 3 samples. This imbalance at the highest level is inherited to individual biomarkers, slide images, and extracted image patches. Hence, the evaluation metrics *Precision*, *Recall*, and *F1 Score* are used to evaluate the proposed method. These are the most commonly used metrics for classification problems with class imbalance. The details of these metrics are provided in Appendix A.

### 5.3.2 Results and Discussion

The evaluation of the proposed classifier is performed at four different levels as follows.

*a) Image patch classification using CNN:* The four trained CNN models are evaluated for their effectiveness in classifying image patches extracted from the slide images of different biomarkers.

*b) Slide level biomarker status prediction:* Biomarker status prediction for individual slide images is evaluated at this level.

*c) Patient level biomarker status prediction:* Every patient sample has 10 slide images per biomarker. Patient level biomarker status prediction considers all these 10 images of each biomarker to predict the patient level biomarker status.

*d) Patient level molecular subtype classification:* At this final level, biomarker predictions are consolidated based on the clinical guidelines to classify each patient sample into one of the four molecular subtypes.

At each level the same four-fold cross validation strategy is used. In this section the results of the various experiments are discussed.

Table 5.4: Result of the biomarker image patch classification by the DenseNet CNN used in the proposed method.

<b>Biomarker</b>	<b>Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
ER	Fold 1	0.9698	0.9673	0.9676
	Fold 2	0.9805	0.9796	0.9796
	Fold 3	0.9727	0.9698	0.9703
	Fold 4	0.9432	0.9435	0.9430
	<b>Average</b>	<b>0.9666</b>	<b>0.9651</b>	<b>0.9651</b>
	<b>Std. Dev.</b>	<b>0.0162</b>	<b>0.0153</b>	<b>0.0156</b>
PR	Fold 1	0.9837	0.9838	0.9837
	Fold 2	0.9897	0.9896	0.9896
	Fold 3	0.9919	0.9915	0.9916
	Fold 4	0.9284	0.9207	0.9179
	<b>Average</b>	<b>0.9734</b>	<b>0.9714</b>	<b>0.9707</b>
	<b>Std. Dev.</b>	<b>0.0302</b>	<b>0.0340</b>	<b>0.0354</b>
Ki67	Fold 1	0.9940	0.9940	0.9940
	Fold 2	0.9975	0.9975	0.9975
	Fold 3	0.9970	0.9970	0.9970
	Fold 4	0.9894	0.9893	0.9893
	<b>Average</b>	<b>0.9945</b>	<b>0.9945</b>	<b>0.9945</b>
	<b>Std. Dev.</b>	<b>0.0037</b>	<b>0.0038</b>	<b>0.0038</b>
HER2	Fold 1	0.9565	0.9551	0.9539
	Fold 2	0.9796	0.9786	0.9785
	Fold 3	0.8155	0.8054	0.7957
	Fold 4	0.8034	0.7548	0.7494
	<b>Average</b>	<b>0.8888</b>	<b>0.8735</b>	<b>0.8694</b>
	<b>Std. Dev.</b>	<b>0.0922</b>	<b>0.1102</b>	<b>0.1138</b>

### 5.3.2.1 Results of Image-patch Classification

Four different CNN models are trained to classify image patches extracted from the biomarker images. These models are evaluated using four-fold cross-validation. Results of the cross-validation of ER, PR, Ki67, and HER2 image-patch classification are shown in Table 5.4. The binary classifiers used in case of ER, PR, and Ki67 classify each image patch into Positive and Negative classes with respect to the biomarker response. HER2 patches are classified into three classes such as Positive, Negative, & Equivocal as followed in clinical procedure. The results of ER, PR, & Ki67 show that the respective CNN models are highly effective in classifying the image patches. The HER2 classifier gives a moderate performance with an average F1 score of 0.8694. The shortage in number of training slide images for the Equivocal class is thought to be the reason for this. The deviation in results across the folds is high for HER2 compared to other three biomarkers.

The performances of the four CNN models on the test set are graphically represented in Figure 5.4. In this, Fold 1 performance of the cross validation is captured for each biomarker. Confusion matrices in Figure 5.4(a, d, g, j) portray the number of test samples in the true classes and the predicted classes. The numbers in the major diagonal show the correct classifications and the rest are misclassifications. These numbers are used for metrics such as Precision, Recall etc. The class imbalance in the test set is visible in these diagrams. Also, the dominating numbers across the major diagonals indicate effectiveness of the CNN models for all four biomarkers. Receiver operating characteristic curve (ROC) is used to represent the effectiveness of a classifier system at various threshold levels. In the ROC curve, true positive rate (TPR), also known as Sensitivity, is plotted against false positive rate (FPR) i.e.,  $1 - \text{Specificity}$ . Area under the ROC curve (AUC) is the numerical value that indicates the effectiveness of a classifier model. AUC value ranges from 0 to 1 where the value 1 indicates a perfect classifier model. Figure 5.4(b,e,h,k) shows the ROC curves plotted for the CNN model of the four biomarkers. The curves and the AUC values represent the effectiveness

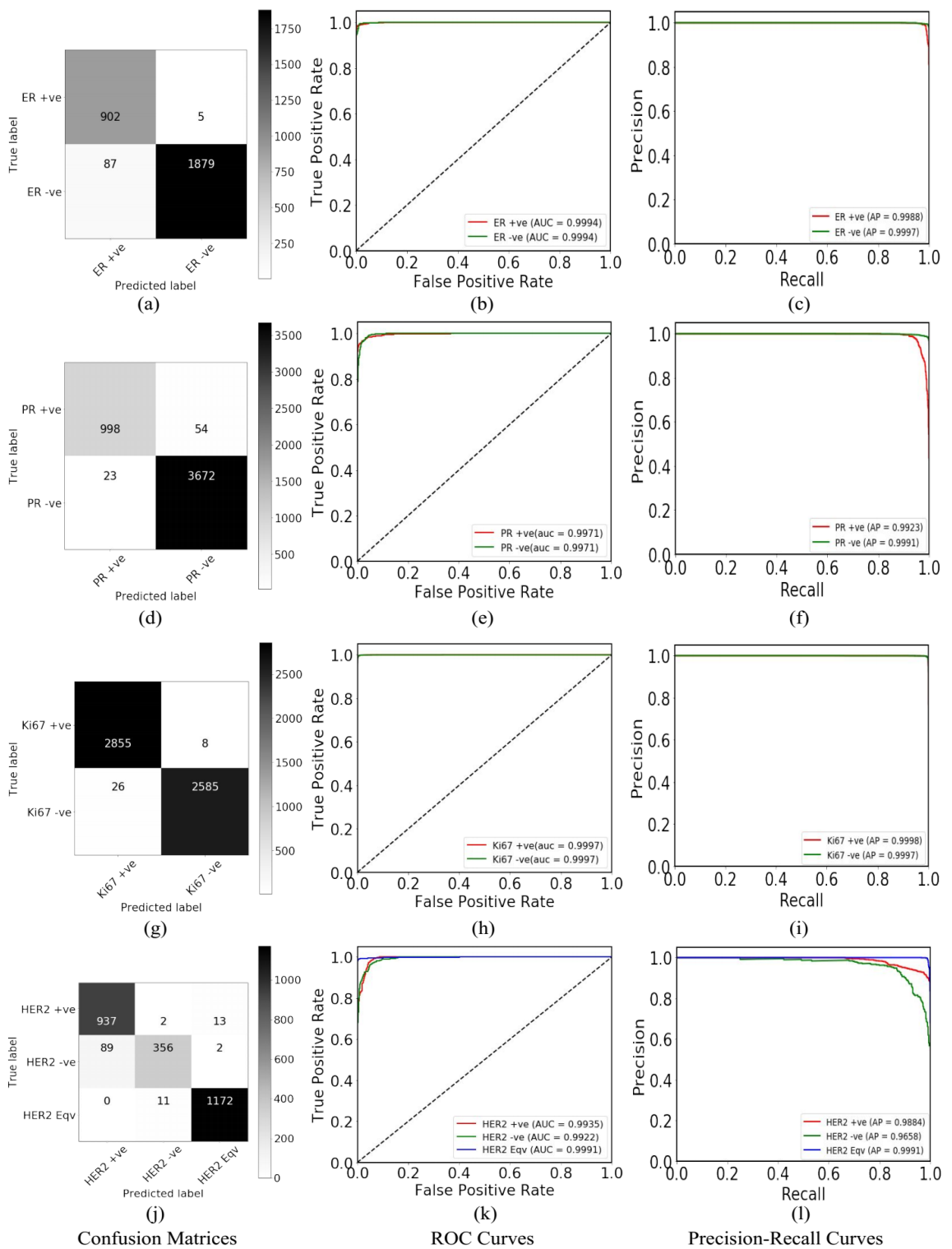


Figure 5.4: Graphs of patch level classifications for Fold 1. Row 1: ER, Row 2: PR, Row 3: Ki67, Row 4: HER2.

of the CNN models by differentiating different classes of IHC image patches. In the datasets with class imbalance, the precision-recall (PR) curve is another method to capture the effectiveness of the classifier. PR curves show a more realistic performance of the model (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015) than ROC curve for imbalanced datasets. Figure 5.4(c,f,i,l) are the corresponding PR curves obtained for the CNN models of the biomarkers ER, PR, Ki67 and HER2. Average precision (AP) value, the weighted mean of precision values obtained at different probability thresholds, summarizes the effectiveness of the PR curve. It can be observed in the graphs that the PR curves portray a more realistic performance of the model than the ROC curves after considering both majority and minority classes in the imbalanced test set. In the case of Ki67, the test set is nearly balanced and hence both ROC and PR curves of the Ki67 model show a similar pattern. Overall high values of AUC and AP indicate that all the CNN models are effective in classifying the IHC biomarker image patches.

### **5.3.2.2 Results of Slide Level Biomarker Status Prediction**

In this set of experiments, the result of image patch classification (i.e., number of patches per class) is processed for each slide image to predict biomarker status at slide level. The prediction algorithm used for each biomarker is explained under the methodology section (Section 5.2.2.3). The slide level prediction performance is evaluated using the ground-truth labels provided for each slide image and the results are summarized in Table 5.5. There are 50 test slide images per biomarker in each fold. Considering all the four folds of cross-validation, the complete set of 200 slide images are evaluated for every biomarker.

The ER and PR slide images that belong to the two target classes (+ve or -ve) are predicted using a threshold for the proportion of positive patches obtained in the patch-level prediction. Biomarker responses of hormone receptors ER and PR are similar in nature and that can be observed in their comparable results as well. Slide level evaluation of ER and PR slide images gives near-perfect results with the average F1



Table 5.5: Result of the slide image classification to target biomarker status classes.

<b>Biomarker</b>	<b>Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
ER	Fold 1	1.0000	1.0000	1.0000
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	1.0000	1.0000	1.0000
	Fold 4	1.0000	1.0000	1.0000
	<b>Average</b>	<b>1.0000</b>	<b>1.0000</b>	<b>1.0000</b>
	<b>Std. Dev.</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>
PR	Fold 1	1.0000	1.0000	1.0000
	Fold 2	0.9473	0.9423	0.9413
	Fold 3	1.0000	1.0000	1.0000
	Fold 4	1.0000	1.0000	1.0000
	<b>Average</b>	<b>0.9868</b>	<b>0.9856</b>	<b>0.9853</b>
	<b>Std. Dev.</b>	<b>0.0264</b>	<b>0.0289</b>	<b>0.0294</b>
Ki67	Fold 1	1.0000	0.8200	0.9011
	Fold 2	1.0000	0.9615	0.9804
	Fold 3	0.9322	0.9322	0.9322
	Fold 4	0.9442	0.9400	0.9358
	<b>Average</b>	<b>0.9691</b>	<b>0.9134</b>	<b>0.9374</b>
	<b>Std. Dev.</b>	<b>0.0360</b>	<b>0.0635</b>	<b>0.0326</b>
HER2	Fold 1	0.9107	0.6875	0.7033
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	0.7731	0.7470	0.7442
	Fold 4	0.8253	0.7857	0.7766
	<b>Average</b>	<b>0.8773</b>	<b>0.8051</b>	<b>0.8060</b>
	<b>Std. Dev.</b>	<b>0.0996</b>	<b>0.1361</b>	<b>0.1327</b>

scores of 0.9658 and 1.00 respectively. The patch level classification by combining all slides in the test set gives results less than 1 for ER and PR (0.9651 for ER and 0.9707 for PR). The consolidation of patch-level classification using the proportion threshold has resulted in improved results for the slide level prediction.

In the case of Ki67, the slide level prediction is more complex. Here the slide images are classified as Ki67 Low/High. In the manual pathology procedure, this is done using a threshold 14% on the proportion of Ki67 +ve nuclei. In the proposed method, this is approximated using the proportion of Ki67 +ve image patches in the slide images. In addition to the base criteria of 14% Ki67+ve patches, experiments are conducted with an additional criterion on the count of +ve patches per slide image. The graph

Table 5.6: Result of applying different threshold criteria for Ki67 slide level status prediction as Ki67 High/Low. (PPP : Positive patch percentage, PPC: Positive patch count)

Criteria	Precision	Recall	F1 Score
PPP: 14%, PPC: Not applied	0.8575	0.8950	0.8550
PPP: 14%, PPC: 24	<b>0.9784</b>	0.8865	0.9226
PPP: 14%, PPC: 20	0.9691	0.9134	<b>0.9374</b>
PPP: 14%, PPC: 18	0.9606	<b>0.9238</b>	0.9370

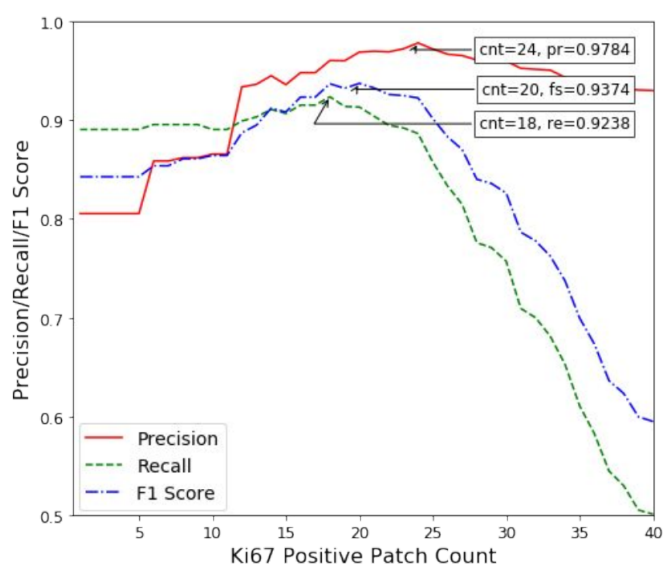


Figure 5.5: Determination of optimal value for minimum positive patch count (cnt) per slide in Ki67 status prediction. Precision (pr) peaked when cnt is kept as 24, Recall (re) peaked for cnt value 18, and F1 score (fs) showed maximum value for cnt value 20.

shown in Figure 5.5 shows the results of these experiments. The slide level prediction results are plotted with varying count threshold for Ki67 +ve patches. It is observed that the precision, recall, and F1 score values are maximum when the minimum Ki67 +ve patch count is kept at 24,18, and 20 respectively. Table 5.6 shows the quantitative results obtained for different criteria applied for Ki67 status prediction. The result obtained using the additional count parameter shows significant improvement over the results

obtained using only the proportion criteria of 14%.

The status prediction of HER2 slide images is also similar to ER and PR except that there are three target classes involved namely positive, and negative, and equivocal. A simple plurality voting among the patches of these three classes is used to predict the slide level biomarker status. In manual evaluation by a pathologist, HER2 equivocal status is assigned when the biomarker is showing neither truly positive nor negative response characteristics. In the deep learning-based automated approach, the same reason can cause more misclassifications. The results of HER2 prediction at patch level as well as slide level is comparatively low due to this reason. In addition, the shortage of image samples in the equivocal class also affected the feature learning by the CNN.

Table 5.7: Result of patient level biomarker status prediction.

<b>Biomarker</b>	<b>Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
ER	Fold 1	1.0000	1.0000	1.0000
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	1.0000	1.0000	1.0000
	Fold 4	1.0000	1.0000	1.0000
	<b>Average</b>	<b>1.0000</b>	<b>1.0000</b>	<b>1.0000</b>
	<b>Std. Dev.</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>
PR	Fold 1	1.0000	1.0000	1.0000
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	1.0000	1.0000	1.0000
	Fold 4	1.0000	1.0000	1.0000
	<b>Average</b>	<b>1.0000</b>	<b>1.0000</b>	<b>1.0000</b>
	<b>Std. Dev.</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>
Ki67	Fold 1	1.0000	0.8000	0.8889
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	0.6400	0.8000	0.7111
	Fold 4	1.0000	1.0000	1.0000
	<b>Average</b>	<b>0.9100</b>	<b>0.9000</b>	<b>0.9000</b>
	<b>Std. Dev.</b>	<b>0.1800</b>	<b>0.1155</b>	<b>0.1364</b>
HER2	Fold 1	0.9000	0.8000	0.8133
	Fold 2	1.0000	1.0000	1.0000
	Fold 3	1.0000	1.0000	1.0000
	Fold 4	1.0000	0.9500	0.9722
	<b>Average</b>	<b>0.9250</b>	<b>0.8500</b>	<b>0.8567</b>
	<b>Std. Dev.</b>	<b>0.0500</b>	<b>0.1000</b>	<b>0.0958</b>

### 5.3.2.3 Patient Level Biomarker Status Prediction

Patient-level evaluation of biomarkers is an extension of slide level evaluation by considering patches from multiple slide images of a patient. In the clinical procedure, normally 10 hotspots in the pathology slides are evaluated to predict patient level biomarker status. In the proposed method also, all the 10 slide images per biomarker of every patient sample are considered to make the consolidated patient level biomarker status prediction. For this, the patches are extracted from all the 10 slide images of a biomarker for a patient and inputted to the CNN for classification. The algorithms and threshold used to predict the status of ER, PR, and HER2 at the patient level are the same as slide level prediction. In the case of Ki67, the minimum patch count criteria used is 200 since 10 Ki67 images are considered now for a patient.

Result of patient level biomarker status prediction is shown in Table 5.7. The results obtained at patient level have remained same or improved for ER, PR, and HER2 whereas the Ki67 results have declined in comparison to slide level predictions. Noticeably, ER and PR predictions have been completely correct for all the patient samples involved in cross-validation. Ki67 and HER2 predictions show scope for further improvements in the automated analysis. High variation in the performance is observed across the cross-validation folds of Ki67 and HER2. In the case of Ki67, the dataset has more borderline slide images for some patients that are labeled based on small differences in the nuclei response. Such samples have high chances of being misclassified. Since at patient-level, the test set has only five patient samples, the cost of one misclassification is high (20%). That leads to high variance in the case of Ki67 performance across the folds. HER2 status assessment is a three-class problem (+ve/-ve/Equivocal) and the borderline case samples are relatively high in this case also. The relatively high variance and lower performance of HER2 status prediction is attributed to the misclassifications of borderline samples. Once the number of patient samples are increased in the dataset, the performance is likely to stabilize with better generalization of the model for both Ki67 and HER2.

Table 5.8: Result of patient-wise molecular subtype classification.

<b>Cross-validation Fold</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Fold1	0.7222	0.6667	0.6667
Fold2	1.0000	1.0000	1.0000
Fold3	1.0000	0.8333	0.9048
Fold4	1.0000	1.0000	1.0000
<b>Average</b>	<b>0.9306</b>	<b>0.8750</b>	<b>0.8929</b>
<b>Std. Dev.</b>	<b>0.1389</b>	<b>0.1596</b>	<b>0.1573</b>

### 5.3.2.4 Molecular Subtype Classification

The patient level molecular subtyping is evaluated in the final set of experiments. The results of these experiments are presented in Table 5.8. This is the culmination of a series of automated processes starting with digitized IHC images of the biomarkers ER, PR, Ki67, and HER2 as the input. Patient level status computed in the previous stage are combined as per the clinical guidelines summarized in Table 5.1 to determine the molecular subtype of the breast cancer for a patient under evaluation. The results obtained are promising with an average precision, recall and F1-score of 0.9306, 0.8750 and 0.8929 respectively. The prediction errors found for biomarkers Ki67 and HER2 (Table 5.7) are reflected in the final subtype classification also. For instance, in the patient level biomarker evaluation, Fold 1 and Fold 3 of Ki67 have given low performance. Fold 1 of HER2 also shows low performance resulting from misclassifications. Since molecular subtyping uses the same folds for evaluation, Fold 1 and Fold 3 of molecular subtyping also have given lower performance. Such performance issues can be addressed by having a large sample size of patients in the dataset that can sufficiently represent all target classes.

### 5.3.2.5 Discussions and Future Scope

The proposed method for automated molecular subtyping is the first ever effort to automate this clinically relevant procedure in breast cancer treatment. The results obtained for the proposed method give a promising outlook for the future development and appli-

cation in clinical practice. The study of the existing literature revealed a few interesting factors. The various methods in the literature assess one or two individual biomarkers that are involved in the molecular subtyping (Table 2.5). These methods mostly use custom datasets that are not available in the public domain. Moreover, there has not been any attempt to take it to the next level of a comprehensive assessment of all the biomarkers to determine the molecular subtype. Unavailability of a patient level dataset covering all the biomarkers is observed as the potential hurdle for this. A dataset of 800 IHC images, collected from 20 patients whose molecular subtypes are known, is used to develop the proposed method. That means, every patient sample has 40 IHC images of dimension  $1920 \times 1440$  covering the four biomarkers. The only limitation observed is in the sample size which is only 20 patients. To increase the sample size to 50 or so, it requires capturing and labeling a total of 1200 more such images. That requires substantial effort from pathologists. A positive observation is that even with the small patient level sample size, the proposed method gives significant performance. The consistency of the results is verified using cross-validation approach.

Despite the robust performance of the proposed method, there are areas that can further improve in future. The use of more recent and better performing deep learning algorithms is one such aspect. Deep learning is a fast-developing domain with newer and better algorithms being developed more frequently. Replacing the DenseNet architecture used in the method with a better performing algorithm can improve results further. Dataset is another factor that can improve the performance. Sample size of certain biomarker response classes is relatively small in the dataset used. HER2 Equivocal class is one such case. It had a negative impact on the performance of HER2 prediction at different levels. If the dataset contains sufficient samples from all the classes, the performance can be improved further. It is also preferable to have public datasets that will help to compare the performance of different methods. The detailed assessment of biomarker response is another area that to be explored. In this method, the focus is on molecular subtyping and only the subjective classes of biomarker responses are considered for it (E.g., ER +ve/-ve, Ki67 low/high etc.). It is also possible to elicit other finer

details from the IHC images such as count or proportion of cells in different subclasses of hormone receptor responses (Weak, Intermediate, Strong etc.) (Allred *et al.*, 1998) that can help oncologists in taking treatment decisions.

## 5.4 Summary

Molecular subtyping classifies breast cancer based on the expression of underlying genetic factors behind the disease. It helps in prognosis and targeted treatment of the disease. A commonly used pathology procedure for molecular subtyping involves IHC analysis of tumor tissues. This manual procedure is tedious and time-consuming. In this chapter, a novel deep learning-based classifier framework of automated molecular subtyping of breast cancer is presented. A combination of traditional image processing and deep learning algorithms are combined in the processing pipeline of the framework. The IHC images of the biomarkers ER, PR, Ki67, and HER2 are independently processed to predict the status of each of these. The results are then combined to determine the molecular subtype of the cancer as per the clinical guidelines. No such system is found in literature currently for automated molecular subtyping based on IHC image analysis. The proposed method is thoroughly evaluated at different levels and the results are found to be highly in concordance with the pathologists' evaluation. The promising results obtained using a small patient sample size is a strong indication of the possibility of achieving the performance level required for the clinical usage of automated methods. Improvement of the results is possible using more advanced CNNs and enhancement of the dataset with more patient samples.





## CHAPTER 6

### CONCLUSIONS

Automated analysis of histopathology images has a great potential to improve the manual procedures in cancer diagnosis and treatment planning in terms of accuracy, affordability, and time required. This thesis focuses on breast cancer histopathology image analysis to develop automated procedures for grading and molecular subtyping of breast cancer. Towards this, deep learning-based automated methods are proposed for breast cancer related pathology procedures namely *mitosis detection*, *nuclear atypia scoring*, and *molecular subtyping*.

The challenges involved in automated mitosis detection are identified through a detailed study of the literature. The study revealed some insightful aspects of existing methods, mainly the evolution of methodology adopted by researchers over the years (Appendix B), their preferences for the choice of basic feature extraction and learning algorithms (Appendix C). The performances of the handcrafted feature-based methods are observed to saturate over time whereas the application of deep learning algorithms is constrained by the class-imbalance problem and dataset sample size. Algorithms trained on single independent datasets are unlikely to perform well on a new dataset since the impact of staining and acquisition settings is high on H & E images, resulting in large variations among images from different datasets. In an attempt to develop generalizable practical solutions based on deep learning algorithms, the proposed method for mitosis detection resorted to merging of datasets from different sources after required preprocessing to normalize the variations. Combining this with patch-level augmentation, the use of an advanced CNN like DenseNet is enabled in the proposed process pipeline. The performance improvement given by this method is an indication to the feasibility and promise of the approach adopted.

Automated nuclear atypia scoring posed an additional set of challenges in terms of the structural complexity of H & E images captured at  $20\times$  magnification, large image dimension, inter-class similarity, and intra-class variations. These were bottlenecks for exploiting the power of deep learning algorithms for this task. The proposed framework for atypia attempted reformulation of the three-class problem of slide level nuclear atypia scoring into a six-class nuclei classification problem. This way the self-learning potential of CNNs is fully utilized in the framework. The aggregation of nuclei-level classification in the post processing stage to predict the atypia score at a slide level showed significant improvement in performance over the state-of-the-art. An important observation from this study is that a combination of conventional image processing techniques and deep learning algorithms like CNNs together can be quite effective for problems that are inherently not suitable for direct application of deep learning algorithms. In addition, a brief analysis of pathologists' disagreement and labeling discrepancies in the dataset are carried out since these problems impact the development of automated methods.

Mitosis detection and atypia scoring are constituent steps of a higher-level task of breast cancer grading whereas the proposed method for automated molecular subtyping is an end-to-end automation of this pathology procedure. IHC images of a biomarker undergo a series of the processing steps in the different stages of the pipeline to give the biomarker status as the output. In this framework also, a preprocessing stage of traditional image processing followed by deep learning and a post processing stage are involved. Applying this pipeline with minimal variations to all the four biomarkers ER, PR, Ki67, and HER2 results in patient-level molecular subtyping of breast cancer. The practical implication of the automation of this task is that the workload of a pathologist can be significantly reduced. In the manual procedure, around 40 hotspots (equivalent to 40 IHC images) in the glass slides need to be analyzed by the pathologist for each patient and literally count the nuclei demonstrating different color expressions. Using an automated method for independent analysis or as an assistive technology, a great deal of this workload can be reduced. However, to reach this stage, more research, fine

tuning, and clinical trials are required. The proposed method is a solid first step in that direction.

In summary, this thesis analyzes three significant procedures in breast cancer pathology i.e., mitosis detection, nuclear atypia scoring, and molecular subtyping to understand the problems of the manual procedures and the limitations of existing automated methods. Deep learning based automated methods are proposed for these procedures to address the observed limitations. The positive results shown by the proposed methods vindicate the potential of the adopted approaches in progressing towards clinically applicable solutions.



# APPENDIX A

## Evaluation Metrics

The evaluation metrics used for various methods presented in the thesis are based on the elements of a confusion matrix that summarizes the performance of classifier systems based on the test output. The four elements of the confusion matrix are True Positive ( $TP$ ), True Negative ( $TN$ ), False Positive ( $FP$ ), and False Negative ( $FN$ ). Descriptions of these elements are given in Table A.1. Definitions of precision, recall, F1 score, and accuracy are shown in Eq. (A.1), Eq. (A.2), Eq. (A.3), and Eq. (A.4) respectively. In the case of multi-class classifiers, the metrics are computed for each class separately and averaged, the weighted average for precision, recall, F1 score and simple average in case of accuracy. Precision is also known as positive predictive value ( $PPV$ ) and recall has alternate names as sensitivity, hit rate, or true positive rate ( $TPR$ ).

Table A.1: Definition of confusion matrix elements used in various evaluation metrics.

Parameter	Description
True Positive ( $TP$ )	Positive sample correctly predicted
True Negative ( $TN$ )	Negative sample correctly predicted
False Positive ( $FP$ )	Negative sample predicted as positive
False Negative ( $FN$ )	Positive sample predicted as negative

$$Precision/PPV = \frac{TP}{TP + FP} \quad (A.1)$$

$$Recall/TPR/Sensitivity = \frac{TP}{TP + FN} \quad (A.2)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (A.3)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (A.4)$$

Apart from these metrics, receiver operating characteristic (ROC) graphs generally used to represent the performance of a classifier system at varying thresholds are also used. ROC curve is a plot of TPR vs. false positive rate (FPR) for each class label. The equation for FPR is given in Eq. (A.5).

$$FPR = \frac{FP}{FP + TN} \quad (A.5)$$

## APPENDIX B

### Methodology Adoption Pattern Over the Years for Automated Mitosis Detection

The mitosis detection methods are broadly grouped into three classes as handcrafted feature-based, deep learning-based and combination of the two. This grouping is based on the patterns observed in the existing methods. In the course of this study, some interesting observations are made about the adoption of these methodologies over the years. Figure B.1 shows the evolution of methodology adoption pattern in every four years since 2008. During 2008-11, mitosis detection methods appeared less in literature and those published used custom datasets that are not available in the public domain.

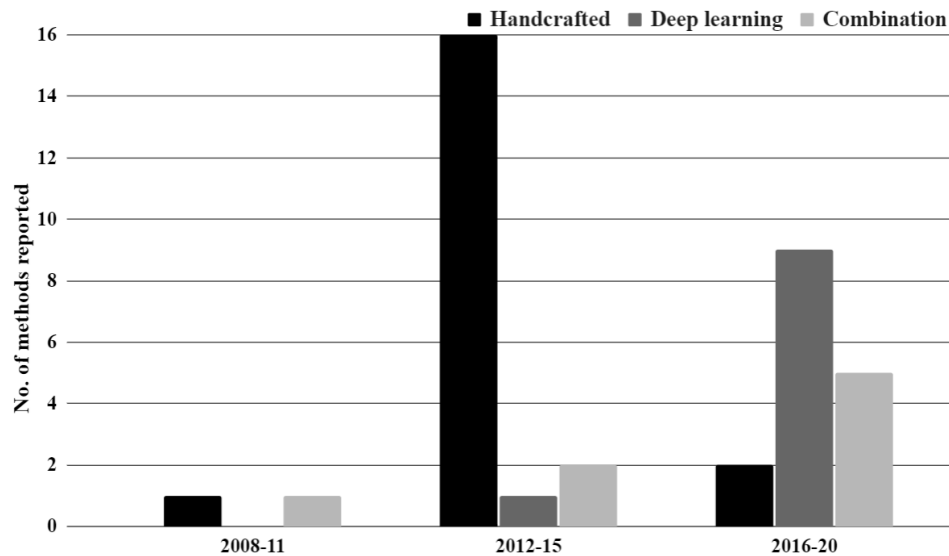


Figure B.1: Methodology adoption trend over the years

During 2012-15, the sudden surge in methods is driven by the open challenges and datasets available in the public domain. Among the methods, handcrafted feature-based methods led the pack. Deep learning was then at an early stage of gaining popularity in medical image processing. Another factor for poor adoption of deep learning during

this period may be attributed to high computing power and large data required for deep learning algorithms. Handcrafted feature-based methods do not require a large amount of data since features are manually extracted, unlike supervised self-learning by deep learning algorithms.

During 2016-19, the trend was reversed. The number of deep learning and combination methods overtook handcrafted feature-based methods by a large margin. The availability of multiple datasets, data augmentation techniques, and availability of computing power has driven this change. However, the result values obtained by deep learning methods were not much different from the results of handcrafted feature-based methods. This is against the trend observed in many other image analysis domains. One possible reason can be that there are very few mitotic cell samples in HPF images compared to non-mitotic cell samples. That means, there is a huge imbalance in the positive and negative samples which affects the process of learning by the deep learning algorithms.



# APPENDIX C

## Adoption Pattern of Standard Algorithms for Automated Mitosis Detection

In the literature study, it is observed that all the reviewed mitosis detection methods use one or more standard algorithms already available, in isolation or combination. This observation provided the motivation to investigate the commonly adopted standard algorithms by researchers in developing their methods. Figure C.1 shows the usage pattern of existing algorithms in various mitosis detection methods reviewed. Following the deep learning trend in recent years, CNNs are used by the maximum number of methods. This is followed by SVM with different kernel functions, random forest, etc. These algorithms are used for different tasks such as feature extraction (CLBP, LOG), segmentation (ACM), classification (RF, DT, CNN), in various phases of mitosis detection.

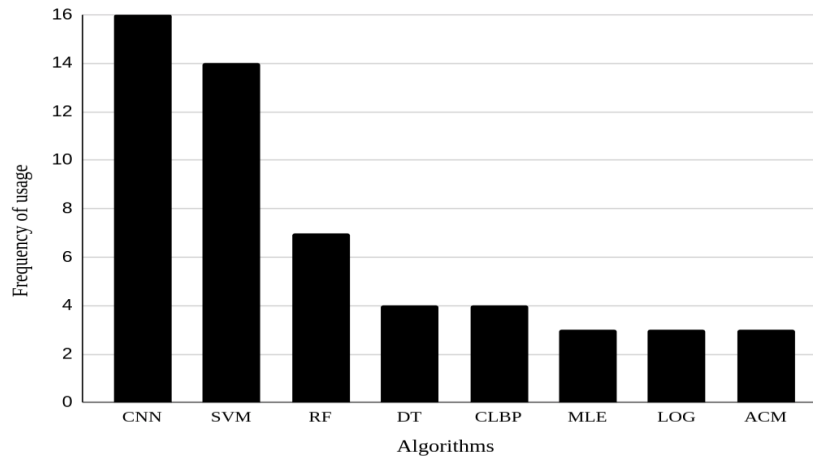


Figure C.1: Usage pattern of standard algorithms in the methods reviewed. (Acronyms: Convolutional neural network (CNN), Support Vector Machine (SVM), Random forest (RF), Decision tree (DT), Completed local binary pattern (CLBP), Maximum-likelihood estimation (MLE), Laplacian of Gaussian (LOG), Active contour model (ACM))



## REFERENCES

- Abubakar, M., J. Figueroa, H. R. Ali, F. Blows, J. Lissowska, C. Caldas, D. F. Easton, M. E. Sherman, M. Garcia-Closas, M. Dowsett, et al.** (2019). Combined quantitative measures of er, pr, her2, and ki67 provide more prognostic information than categorical combinations in luminal breast cancer. *Modern Pathology*, **32**(9), 1244–1256. 26
- Abubakar, M., W. J. Howat, F. Daley, L. Zabaglo, L.-A. McDuffus, F. Blows, P. Coulson, H. Raza Ali, J. Benitez, R. Milne, et al.** (2016). High-throughput automated scoring of ki67 in breast cancer tissue microarrays from the breast cancer association consortium. *The Journal of Pathology: Clinical Research*, **2**(3), 138–153. 29, 30
- Al-Kofahi, Y., W. Lassoued, W. Lee, and B. Roysam** (2009). Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, **57**(4), 841–852. 14, 41, 42, 69, 90
- Al-Thoubaity, F. K.** (2020). Molecular classification of breast cancer: A retrospective cohort study. *Annals of Medicine and Surgery*, **49**, 44–48. 2, 5, 25
- Albarqouni, S., C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab** (2016). Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, **35**(5), 1313–1321. 20
- Allred, D., J. M. Harvey, M. Berardo, and G. M. Clark** (1998). Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, **11**(2), 155–168. 26, 103, 117
- Amin, J., M. Sharif, N. Gul, M. Raza, M. A. Anjum, M. W. Nisar, and S. A. C. Bukhari** (2020). Brain tumor detection by using stacked autoencoders in deep learning. *Journal of Medical Systems*, **44**(2), 32. 19
- Anwar, S. M., M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan** (2018). Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, **42**(11), 226. 19
- Ayad, E., M. Mansy, D. Elwi, M. Salem, M. Salama, and K. Kayser** (2015). Comparative study between quantitative digital image analysis and fluorescence in situ hybridization of breast cancer equivocal human epidermal growth factor receptors 2 score 2+ cases. *Journal of pathology informatics*, **6**. 27

- Basavanhally, A., E. Yu, J. Xu, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi**, Incorporating domain knowledge for tubule detection in breast histopathology using o'callaghan neighborhoods. *In Medical Imaging 2011: Computer-Aided Diagnosis*, volume 7963. International Society for Optics and Photonics, 2011. 3
- Beevi, K. S., M. S. Nair, and G. Bindu**, Detection of mitotic nuclei in breast histopathology images using localized acm and random kitchen sink based classifier. *In 2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016. 18
- Beevi, K. S., M. S. Nair, and G. Bindu** (2017). A multi-classifier system for automatic mitosis detection in breast histopathology images using deep belief networks. *IEEE journal of translational engineering in health and medicine*, **5**, 1–11. 21
- Beevi, K. S., M. S. Nair, and G. Bindu** (2019). Automatic mitosis detection in breast histopathology images using convolutional neural network based deep transfer learning. *Biocybernetics and Biomedical Engineering*, **39**(1), 214–223. 21, 22
- Bloom, H. and W. Richardson** (1957). Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, **11**(3), 359. 3, 14
- Boykov, Y. and V. Kolmogorov** (2004a). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence*, **26**(9), 1124–1137. 41, 42
- Boykov, Y. and V. Kolmogorov** (2004b). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence*, **26**(9), 1124–1137. 69
- Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal** (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, **68**(6), 394–424. 36
- Bryne, M., H. S. Koppang, R. Lilleng, T. Stene, G. Bang, and E. Dabelsteen** (1989). New malignancy grading is a better prognostic indicator than broders' grading in oral squamous cell carcinomas. *Journal of Oral Pathology & Medicine*, **18**(8), 432–437. 2
- Buda, M., A. Maki, and M. A. Mazurowski** (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, **106**, 249–259. 48
- Cai, D., X. Sun, N. Zhou, X. Han, and J. Yao**, Efficient mitosis detection in breast cancer histology images by renn. *In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019. 20, 21, 54, 55

- Chan, T. F.** and **L. A. Vese** (2001). Active contours without edges. *IEEE Transactions on image processing*, **10**(2), 266–277. 18
- Chang, R.-F., H.-H. Chen, Y.-C. Chang, C.-S. Huang, J.-H. Chen,** and **C.-M. Lo** (2016). Quantification of breast tumor heterogeneity for er status, her2 status, and tn molecular subtype evaluation on dce-mri. *Magnetic resonance imaging*, **34**(6), 809–819. 26
- Chaudhury, B., M. Zhou, D. B. Goldgof, L. O. Hall, R. A. Gatenby, R. J. Gillies,** and **J. S. Drukteinis**, Using features from tumor subregions of breast dce-mri for estrogen receptor status prediction. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014. 26
- Chawla, N. V.** (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875–886. 48
- Chen, H., Q. Dou, X. Wang, J. Qin,** and **P. A. Heng**, Mitosis detection in breast cancer histology images via deep cascaded networks. In *Thirtieth AAAI Conference on Artificial Intelligence*. 2016a. 20, 54, 55
- Chen, H., X. Wang,** and **P. A. Heng**, Automated mitosis detection with deep regression networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016b. 20
- Chung, Y. R., M. H. Jang, S. Y. Park, G. Gong, W.-H. Jung, K. B. P. K.-. S. Group,** et al. (2016). Interobserver variability of ki-67 measurement in breast cancer. *Journal of pathology and translational medicine*, **50**(2), 129. 5
- Cireşan, D. C., A. Giusti, L. M. Gambardella,** and **J. Schmidhuber**, Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*. Springer, 2013. 19, 20
- Cohen, J.** (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46. 88
- Collisson, E. A., P. Bailey, D. K. Chang,** and **A. V. Biankin** (2019). Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology*, **16**(4), 207–220. 2
- Cosatto, E., M. Miller, H. P. Graf,** and **J. S. Meyer**, Grading nuclear pleomorphism on histological micrographs. In *2008 19th International Conference on Pattern Recognition*. IEEE, 2008. 15
- Dabbs, D. J.,** *Diagnostic Immunohistochemistry E-Book: Theranostic and Genomic Applications*. Elsevier Health Sciences, 2017. 5

- Dalle, J.-R., W. K. Leow, D. Racoceanu, A. E. Tutac, and T. C. Putti**, Automatic breast cancer grading of histopathological images. *In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008. 7, 15
- Dalle, J.-R., H. Li, C.-H. Huang, W. K. Leow, D. Racoceanu, and T. C. Putti**, Nuclear pleomorphism scoring by selective cell nuclei detection. *In WACV*. 2009. 15
- Dalton, L. W., S. E. Pinder, C. E. Elston, I. O. Ellis, D. L. Page, W. D. Dupont, and R. W. Blamey** (2000). Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Modern pathology*, **13**(7), 730–735. 87
- Das, A., M. S. Nair, and D. S. Peter** (2020a). Batch mode active learning on the riemannian manifold for automated scoring of nuclear pleomorphism in breast cancer. *Artificial Intelligence in Medicine*, **103**, 101805. 24, 84
- Das, A., M. S. Nair, and S. D. Peter** (2018). Sparse representation over learned dictionaries on the riemannian manifold for automated grading of nuclear pleomorphism in breast cancer. *IEEE Transactions on Image Processing*, **28**(3), 1248–1260. 24, 32, 84, 85
- Das, A., M. S. Nair, and S. D. Peter** (2019). Kernel-based fisher discriminant analysis on the riemannian manifold for nuclear atypia scoring of breast cancer. *Biocybernetics and Biomedical Engineering*, **39**(3), 728–741. 24, 84, 85, 87
- Das, A., M. S. Nair, and S. D. Peter** (2020b). Computer-aided histopathological image analysis techniques for automated nuclear atypia scoring of breast cancer: a review. *Journal of Digital Imaging*, 1–31. 3, 60
- Das, D. K. and P. K. Dutta** (2019). Efficient automated detection of mitotic cells from breast histological images using deep convolution neural network with wavelet decomposed patches. *Computers in biology and medicine*, **104**, 29–42. 20, 21, 54, 55
- Davis, J. and M. Goadrich**, The relationship between precision-recall and roc curves. *In Proceedings of the 23rd international conference on Machine learning*. 2006. 51, 110
- Doyle, S., S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski**, Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. *In 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2008. 7, 15
- Dunne, B. and J. Going** (2001). Scoring nuclear pleomorphism in breast cancer. *Histopathology*, **39**(3), 259–265. 14, 60, 61
- Eaden, J., K. Abrams, H. McKay, H. Denley, and J. Mayberry** (2001). Inter-observer variation between general and specialist gastrointestinal pathologists when

grading dysplasia in ulcerative colitis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **194**(2), 152–157. 14

**Einstein, A. J., H.-S. Wu, M. Sanchez, and J. Gil** (1998). Fractal characterization of chromatin appearance for diagnosis in breast cytology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **185**(4), 366–381. 15

**Eliyatkın, N., E. Yalçın, B. Zengel, S. Aktaş, and E. Vardar** (2015). Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way. *The journal of breast health*, **11**(2), 59. 15, 94

**Elmore, J. G., G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, et al.** (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, **313**(11), 1122–1132. 14

**Elston, C. W. and I. O. Ellis** (2002). Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. cw elston & io ellis. *histopathology* 1991; 19; 403–410: Author commentary. *Histopathology*, **41**(3a), 151–151. 3, 14, 59

**Epstein, J. I., L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey** (2016). The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, **40**(2), 244–252. 2

**Fan, J.** (1998). Notes on poisson distribution-based minimum error thresholding. *Pattern Recognition Letters*, **19**(5-6), 425–431. 41

**Frierson Jr, H. F., R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, and D. C. Wilbur** (1995). Interobserver reproducibility of the nottingham modification of the bloom and richardson histologic grading scheme for infiltrating ductal carcinoma. *American journal of clinical pathology*, **103**(2), 195–198. 60, 87

**Fuchs, T. J. and J. M. Buhmann** (2011). Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, **35**(7-8), 515–530. 15, 36

**Fuhrman, S. A., L. C. Lasky, and C. Limas** (1982). Prognostic significance of morphologic parameters in renal cell carcinoma. *The American journal of surgical pathology*, **6**(7), 655–664. 2

**Gandomkar, Z., P. C. Brennan, and C. Mello-Thoms** (2019). Computer-assisted nuclear atypia scoring of breast cancer: a preliminary study. *Journal of digital imaging*, **32**(5), 702–712. 24, 60

- Gavrielides, M. A., B. D. Gallas, P. Lenz, A. Badano, and S. M. Hewitt** (2011). Observer variability in the interpretation of her2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. *Archives of pathology & laboratory medicine*, **135**(2), 233–242. 5
- Gerdes, J., L. Li, C. Schlueter, M. Duchrow, C. Wohlenberg, C. Gerlach, I. Stahmer, S. Kloth, E. Brandt, and H.-D. Flad** (1991). Immunobiochemical and molecular biologic characterization of the cell proliferation-associated nuclear antigen that is defined by monoclonal antibody ki-67. *The American journal of pathology*, **138**(4), 867–29
- Goldhirsch, A., E. P. Winer, A. Coates, R. Gelber, M. Piccart-Gebhart, B. Thürlimann, H.-J. Senn, K. S. Albain, F. André, J. Bergh, et al.** (2013). Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013. *Annals of oncology*, **24**(9), 2206–2223. 5
- Greenough, R. B.** (1925). Varying degrees of malignancy in cancer of the breast. *The Journal of Cancer Research*, **9**(4), 453–463. 14
- Gu, J., Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al.** (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, **77**, 354–377. 44
- Guinney, J., R. Dienstmann, X. Wang, A. De Reynies, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, et al.** (2015). The consensus molecular subtypes of colorectal cancer. *Nature medicine*, **21**(11), 1350–1356. 2
- Gurovich, Y., Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, et al.** (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, **25**(1), 60–64. 19
- Hall, B. H., M. Ianosi-Irimie, P. Javidian, W. Chen, S. Ganesan, and D. J. Foran** (2008). Computer-assisted assessment of the human epidermal growth factor receptor 2 immunohistochemical assay in imaged histologic sections using a membrane isolation algorithm and quantitative analysis of positive controls. *BMC Medical Imaging*, **8**(1), 1–13. 28
- Harbeck, N. and M. Gnant** (2016). Breast cancer. *Lancet (London, England)*, **389**(10074), 1134–1150. 3
- Harvey, J. M., G. M. Clark, C. K. Osborne, D. C. Allred, et al.** (1999). Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of clinical oncology*, **17**(5), 1474–1481. 103



- He, K., X. Zhang, S. Ren, and J. Sun**, Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. 49, 73
- Henson, D. E., L. Ries, L. S. Freedman, and M. Carriaga** (1991). Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. the basis for a prognostic index. *Cancer*, **68**(10), 2142–2149. 14
- Higgins, C.** (2015). Applications and challenges of digital pathology and whole slide imaging. *Biotechnic & Histochemistry*, **90**(5), 341–347. 6
- Houssein, E. H., M. M. Emam, A. A. Ali, and P. N. Suganthan** (2020). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, 114161. 23
- Huang, C.-H. and H.-K. Lee**, Automated mitosis detection based on exclusive independent component analysis. *In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012. 17
- Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger**, Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. 44, 49, 73, 90, 98
- ICPR** (2012). Mitosis - dataset. [http://ludo17.free.fr/mitos\\_2012/dataset.html](http://ludo17.free.fr/mitos_2012/dataset.html). 7, 46, 55
- ICPR** (2014). Mitos-atypia - dataset. <https://mitos-atypia-14.grand-challenge.org/Dataset/>. 16, 38, 40, 45, 46, 55, 61, 78
- Irshad, H.** (2013). Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *Journal of pathology informatics*, **4**. 17, 18
- Irshad, H., A. Gouillard, L. Roux, and D. Racoceanu** (2014a). Multispectral band selection and spatial characterization: Application to mitosis detection in breast cancer histopathology. *Computerized Medical Imaging and Graphics*, **38**(5), 390–402. 17, 18
- Irshad, H., A. Gouillard, L. Roux, and D. Racoceanu**, Spectral band selection for mitosis detection in histopathology. *In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014b. 17, 18
- Irshad, H., S. Jalali, L. Roux, D. Racoceanu, L. J. Hwee, G. Le Naour, and F. Capron** (2013). Automated mitosis detection using texture, sift features and hmax biologically inspired approach. *Journal of pathology informatics*, **4**(Suppl). 17, 18
- Jamaluddin, M. F., M. F. Fauzi, F. S. Abas, J. T. Lee, S. Y. Khor, K. H. Teoh, and L. M. Looi**, Cell classification in er-stained whole slide breast cancer images using convolutional neural network. *In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018. 26, 30

- Johnson, J. M.** and **T. M. Khoshgoftaar** (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, **6**(1), 27. 56
- Kadam, V. J., S. M. Jadhav,** and **K. Vijayakumar** (2019). Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. *Journal of medical systems*, **43**(8), 263. 19
- Kaman, E., A. Smeulders, P. Verbeek, I. Young,** and **J. Baak** (1984). Image processing for mitoses in sections of breast cancer: A feasibility study. *Cytometry: The Journal of the International Society for Analytical Cytology*, **5**(3), 244–249. 15
- Khan, A. M., H. El-Daly,** and **N. M. Rajpoot**, A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. *In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012. 17
- Khan, A. M., N. Rajpoot, D. Treanor,** and **D. Magee** (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, **61**(6), 1729–1738. 39, 67
- Khan, A. M., K. Sirinukunwattana,** and **N. Rajpoot** (2015). A global covariance descriptor for nuclear atypia scoring in breast histopathology images. *IEEE journal of biomedical and health informatics*, **19**(5), 1637–1647. 22, 24, 84, 85, 87
- Kingma, D. P.** and **J. Ba** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 74
- Konsti, J., M. Lundin, H. Joensuu, T. Lehtimäki, H. Sihto, K. Holli, T. Turpeenniemi-Hujanen, V. Kataja, L. Sailas, J. Isola, et al.** (2011). Development and evaluation of a virtual microscopy application for automated assessment of ki-67 expression in breast cancer. *BMC clinical pathology*, **11**(1), 1–11. 30, 31
- Kristiansen, G.** (2018). Next-generation nuclear morphology to grade solid tumours. *The Lancet Oncology*, **19**(3), 275–277. 59
- Kronqvist, P., T. Kuopio,** and **Y. Collan** (1998). Morphometric grading of invasive ductal breast cancer. i. thresholds for nuclear grade. *British journal of cancer*, **78**(6), 800–805. 15
- Lakshmi, S., D. Vijayaseenan, D. S. Sumam, S. Sreeram,** and **P. K. Suresh**, An integrated deep learning approach towards automatic evaluation of ki-67 labeling index. *In TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019. 30, 31
- Li, C., X. Wang, W. Liu,** and **L. J. Latecki** (2018). Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical image analysis*, **45**, 121–133. 20, 54

- Li, X.** and **K. N. Plataniotis** (2015). A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. *IEEE Transactions on Biomedical Engineering*, **62**(7), 1862–1873. 39, 68
- Liao, P.-S., T.-S. Chen, P.-C. Chung, et al.** (2001). A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, **17**(5), 713–727. 100
- Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. I. Sánchez** (2017). A survey on deep learning in medical image analysis. *arXiv preprint arXiv:1702.05747*. 19, 32, 95, 98
- Lloyd, M. C., P. Allam-Nandyala, C. N. Purohit, N. Burke, D. Coppola, and M. M. Bui** (2010). Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: How reliable is it? *Journal of pathology informatics*, **1**. 27
- Lu, C., M. Ji, Z. Ma, and M. Mandal** (2015). Automated image analysis of nuclear atypia in high-power field histopathological image. *journal of microscopy*, **258**(3), 233–240. 22, 24, 84, 85, 87
- Lu, C.** and **M. Mandal** (2013). Toward automatic mitotic cell detection and segmentation in multispectral histopathological images. *IEEE Journal of Biomedical and Health Informatics*, **18**(2), 594–605. 17, 18
- Malhotra, G. K., X. Zhao, H. Band, and V. Band** (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer biology & therapy*, **10**(10), 955–960. 93
- Malon, C., E. Brachtel, E. Cosatto, H. P. Graf, A. Kurata, M. Kuroda, J. S. Meyer, A. Saito, S. Wu, and Y. Yagi** (2012). Mitotic figure recognition: Agreement among pathologists and computerized detector. *Analytical Cellular Pathology*, **35**(2), 97–100. 2, 4, 8, 14, 17
- Malon, C., M. Miller, H. C. Burger, E. Cosatto, and H. P. Graf**, Identifying histological elements with convolutional neural networks. *In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. 2008. 21
- Malon, C. D.** and **E. Cosatto** (2013). Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of pathology informatics*, **4**. 21
- Maqlin, P., R. Thamburaj, J. J. Mammen, and M. T. Manipadam**, Automated nuclear pleomorphism scoring in breast cancer histopathology images using deep neural networks. *In International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2015. 23

- Mathew, T., J. R. Kini, and J. Rajan** (2020). Computational methods for automated mitosis detection in histopathology images: A review. *Biocybernetics and Biomedical Engineering*, **41**(1), 64–82. 7, 36
- Matsumoto, M. and T. Nishimura** (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **8**(1), 3–30. 79
- McCarty, K. S., E. Szabo, J. L. Flowers, E. B. Cox, G. S. Leight, L. Miller, J. Konrath, J. T. Soper, D. A. Budwit, W. T. Creasman, et al.** (1986). Use of a monoclonal anti-estrogen receptor antibody in the immunohistochemical evaluation of human tumors. *Cancer research*, **46**(8 Supplement), 4244s–4248s. 26, 27
- Mofidi, R., R. Walsh, P. Ridgway, T. Crotty, E. McDermott, T. Keaveny, M. Duffy, A. Hill, and N. O’Higgins** (2003). Objective measurement of breast cancer oestrogen receptor status through digital image analysis. *European Journal of Surgical Oncology (EJSO)*, **29**(1), 20–24. 25
- Moncayo, R., D. Romo-Bucheli, and E. Romero**, A grading strategy for nuclear pleomorphism in histopathological breast cancer images using a bag of features (bof). *In Iberoamerican Congress on Pattern Recognition*. Springer, 2015. 23
- Mouelhi, A., M. Sayadi, and F. Fnaiech**, A novel morphological segmentation method for evaluating estrogen receptors’ status in breast tissue images. *In 2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 2014. 26, 30
- Mulrane, L., E. Rexhepaj, S. Penney, J. J. Callanan, and W. M. Gallagher** (2008). Automated image analysis in histopathology: a valuable tool in medical diagnostics. *Expert review of molecular diagnostics*, **8**(6), 707–725. 6
- Naik, S., S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski**, Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. *In 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2008. 7, 15
- Nair, V. and G. E. Hinton**, Rectified linear units improve restricted boltzmann machines. *In ICML*. 2010. 74
- Nateghi, R., H. Danyali, and M. S. Helfroush** (2017). Maximized inter-class weighted mean for fast and accurate mitosis cells detection in breast cancer histopathology images. *Journal of medical systems*, **41**(9), 146. 17
- Nateghi, R., H. Danyali, M. Sadegh Helfroush, and A. Tashk** (2014). Intelligent cad system for automatic detection of mitotic cells from breast cancer histology slide images based on teaching-learning-based optimization. *Computational Biology Journal*. 17

- Niazi, M. K. K., A. V. Parwani, and M. N. Gurcan** (2019). Digital pathology and artificial intelligence. *The lancet oncology*, **20**(5), e253–e261. 9, 95
- Niazi, M. K. K., M. M. Yearsley, X. Zhou, W. L. Frankel, and M. N. Gurcan** (2014). Perceptual clustering for automatic hotspot detection from ki-67-stained neuroendocrine tumour images. *Journal of microscopy*, **256**(3), 213–225. 30, 31
- Nicholson, A., B. Addis, H. Bharucha, C. Clelland, B. Corrin, A. Gibbs, P. Hasleton, K. Kerr, N. Ibrahim, S. Stewart, et al.** (2004). Inter-observer variation between pathologists in diffuse parenchymal lung disease. *Thorax*, **59**(6), 500–505. 14
- Onder, D., S. Zengin, and S. Sarioglu** (2014). A review on color normalization and color deconvolution methods in histopathology. *Applied Immunohistochemistry & Molecular Morphology*, **22**(10), 713–719. 67
- Oscanoa, J., F. Doimi, R. Dyer, J. Araujo, J. Pinto, and B. Castaneda**, Automated segmentation and classification of cell nuclei in immunohistochemical breast cancer images with estrogen receptor marker. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016. 26, 30
- Pang, T., J. H. D. Wong, W. L. Ng, and C. S. Chan** (2020). Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Systems with Applications*, 113501. 32
- Paul, A. and D. P. Mukherjee** (2015). Mitosis detection for invasive breast cancer grading in histopathological images. *IEEE transactions on image processing*, **24**(11), 4041–4054. 17, 18, 36
- Perez, E. A., J. Cortés, A. M. Gonzalez-Angulo, and J. M. Bartlett** (2014). Her2 testing: current status and future directions. *Cancer treatment reviews*, **40**(2), 276–284. 27
- Pienta, K. J. and D. S. Coffey** (1991). Correlation of nuclear morphometry with progression of breast cancer. *Cancer*, **68**(9), 2012–2016. 14, 59
- Pitkäaho, T., T. M. Lehtimäki, J. McDonald, T. J. Naughton, et al.**, Classifying her2 breast cancer cell samples using deep learning. In *Proc. Irish Mach. Vis. Image Process. Conf.*. 2016. 28, 30
- Prat, A., E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and M. Muñoz** (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, **24**, S26–S35. 93
- Raimondo, F., M. A. Gavrielides, G. Karayannopoulou, K. Lyroudia, I. Pitas, and I. Kostopoulos** (2005). Automated evaluation of her-2/neu status in breast tissue from fluorescent in situ hybridization images. *IEEE Transactions on Image Processing*, **14**(9), 1288–1299. 27

- Ramzan, F., M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood** (2020). A deep learning approach for automated diagnosis and multi-class classification of alzheimer’s disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, **44**(2), 37. 19
- Reinhard, E., M. Adhikhmin, B. Gooch, and P. Shirley** (2001). Color transfer between images. *IEEE Computer graphics and applications*, **21**(5), 34–41. 39, 55, 67, 68, 90
- Rexhepaj, E., D. J. Brennan, P. Holloway, E. W. Kay, A. H. McCann, G. Landberg, M. J. Duffy, K. Jirstrom, and W. M. Gallagher** (2008). Novel image analysis approach for quantifying expression of nuclear proteins assessed by immunohistochemistry: application to measurement of oestrogen and progesterone receptor levels in breast cancer. *Breast Cancer Research*, **10**(5), 1–10. 26, 30
- Rezaeilouyeh, H., A. Mollahosseini, and M. H. Mahoor** (2016). Microscopic medical image classification framework via deep learning and shearlet transform. *Journal of Medical Imaging*, **3**(4), 044501. 24, 84, 85, 87
- Robbins, P., S. Pinder, N. De Klerk, H. Dawkins, J. Harvey, G. Sterrett, I. Ellis, and C. Elston** (1995). Histological grading of breast carcinomas: a study of interobserver agreement. *Human pathology*, **26**(8), 873–879. 2, 4, 14, 87
- Romo-Bucheli, D., A. Janowczyk, H. Gilmore, E. Romero, and A. Madabhushi** (2017). A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers. *Cytometry Part A*, **91**(6), 566–573. 20
- Roullier, V., O. L  zoray, V.-T. Ta, and A. Elmoataz** (2011). Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics*, **35**(7-8), 603–615. 17
- Roux, L., D. Racoceanu, N. Lom  nie, M. Kulikova, H. Irshad, J. Klossa, F. Capron, C. Genestie, G. Le Naour, and M. N. Gurcan** (2013). Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, **4**. 7, 8
- Saha, M., I. Arun, R. Ahmed, S. Chatterjee, and C. Chakraborty** (2020). Hscorenet: A deep network for estrogen and progesterone scoring using breast ihc images. *Pattern Recognition*, **102**, 107200. 26, 30
- Saha, M. and C. Chakraborty** (2018). Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, **27**(5), 2189–2200. 28, 30

- Saha, M., C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee** (2017). An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. *Scientific reports*, **7**(1), 1–14. 30, 31
- Saha, M., C. Chakraborty, and D. Racoceanu** (2018). Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*, **64**, 29–40. 21, 22
- Saito, T. and M. Rehmsmeier** (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, **10**(3), e0118432. 51, 110
- Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen**, Mobilenetv2: Inverted residuals and linear bottlenecks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. 73
- Sawair, F. A., C. R. Irwin, D. J. Gordon, A. G. Leonard, M. Stephenson, and S. S. Napier** (2003). Invasive front grading: reliability and usefulness in the management of oral squamous cell carcinoma. *Journal of oral pathology & medicine*, **32**(1), 1–9. 2
- Shi, P., J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen** (2016). Automated ki-67 quantification of immunohistochemical staining image of human nasopharyngeal carcinoma xenografts. *Scientific reports*, **6**(1), 1–9. 29, 30
- Shorten, C. and T. M. Khoshgoftaar** (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, **6**(1), 60. 44, 56
- Simonyan, K. and A. Zisserman** (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 49, 73
- Skaland, I., I. Øvestad, E. A. Janssen, J. Klos, K. H. Kjellevoid, T. Helliesen, and J. Baak** (2008). Comparing subjective and digital image analysis her2/neu expression scores with conventional and modified fish scores in breast cancer. *Journal of clinical pathology*, **61**(1), 68–71. 28, 30
- Sommer, C., L. Fiaschi, F. A. Hamprecht, and D. W. Gerlich**, Learning-based mitotic cell detection in histopathological images. *In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012. 17
- Srinidhi, C. L., O. Ciga, and A. L. Martel** (2020). Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, **67**, 101813. 6, 44
- Stierer, M., H. Rosen, and R. Weber** (1991). Nuclear pleomorphism, a strong prognostic factor in axillary node-negative small invasive breast cancer. *Breast cancer research and treatment*, **20**(2), 109–116. 59
- SU** (2001). Stanford university tma database. <https://tma.im/>. 26

- Sung, H., J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray** (2020). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, **71**(3), 209–249. 1, 8
- Swiderska, Z., A. Korzynska, T. Markiewicz, M. Lorent, J. Zak, A. Wesolowska, L. Roszkowiak, J. Slodkowska, and B. Grala** (2015). Comparison of the manual, semiautomatic, and automatic selection and leveling of hot spots in whole slide images for ki-67 quantification in meningiomas. *Analytical cellular pathology*, **2015**. 29
- Tariq, M., S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi** (2020). Medical image based breast cancer diagnosis: State of the art and future directions. *Expert Systems with Applications*, 114095. 23
- Tashk, A., M. S. Helfroush, H. Danyali, and M. Akbarzadeh**, An automatic mitosis detection method for breast cancer histopathology slide images based on objective and pixel-wise textural features classification. *In The 5th conference on information and knowledge technology*. IEEE, 2013. 17, 18
- Tashk, A., M. S. Helfroush, H. Danyali, and M. Akbarzadeh-Jahromi** (2015). Automatic detection of breast cancer mitotic cells based on the combination of textural, statistical and innovative mathematical features. *Applied Mathematical Modelling*, **39**(20), 6165–6182. 17
- Tek, F. B.** (2013). Mitosis detection using generic features and an ensemble of cascade adaboosts. *Journal of pathology informatics*, **4**. 17, 18
- Thakur, N., H. Yoon, and Y. Chong** (2020). Current trends of artificial intelligence for colorectal cancer pathology image analysis: A systematic review. *Cancers*, **12**(7), 1884. 6
- Ting, F. F., Y. J. Tan, and K. S. Sim** (2019). Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications*, **120**, 103–115. 23
- Tsang, J. and G. M. Tse** (2020). Molecular classification of breast cancer. *Advances in anatomic pathology*, **27**(1), 27–35. 93
- Tuominen, V. J., S. Ruotoistenmäki, A. Viitanen, M. Jumppanen, and J. Isola** (2010). Immunoratio: a publicly available web application for quantitative image analysis of estrogen receptor (er), progesterone receptor (pr), and ki-67. *Breast cancer research*, **12**(4), 1–12. 25, 30
- Tuominen, V. J., T. T. Tolonen, and J. Isola** (2012). Immunomembrane: a publicly available web application for digital image analysis of her2 immunohistochemistry. *Histopathology*, **60**(5), 758–767. 28, 30



- Vahadane, A., T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab** (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, **35**(8), 1962–1971. 39, 68
- Vandenbergh, M. E., M. L. Scott, P. W. Scorer, M. Söderberg, D. Balcerzak, and C. Barker** (2017). Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer. *Scientific reports*, **7**(1), 1–11. 28, 30
- Veta, M., P. J. van Diest, and J. P. Pluim**, Detecting mitotic figures in breast cancer histopathology images. In *Medical Imaging 2013: Digital Pathology*, volume 8676. International Society for Optics and Photonics, 2013. 17, 18
- Vijayashree, R., P. Aruthra, and K. R. Rao** (2015). A comparison of manual and automated methods of quantitation of oestrogen/progesterone receptor expression in breast carcinoma. *Journal of clinical and diagnostic research: JCDR*, **9**(3), EC01. 25
- Wahab, N., A. Khan, and Y. S. Lee** (2017). Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. *Computers in biology and medicine*, **85**, 86–97. 20, 21
- Wan, T., J. Cao, J. Chen, and Z. Qin** (2017). Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing*, **229**, 34–44. 23
- Wang, H., C. Jiang, K. Bao, and C. Xu** (2019). Recognition and clinical diagnosis of cervical cancer cells based on our improved lightweight deep network for pathological image. *Journal of medical systems*, **43**(9), 301. 19
- Wang, H., A. C. Roa, A. N. Basavanahally, H. L. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi** (2014). Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, **1**(3), 034003. 21, 22
- Wang, S. and X. Yao** (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42**(4), 1119–1130. 56
- Wardle, J., K. Robb, S. Vernon, and J. Waller** (2015). Screening for prevention and early diagnosis of cancer. *American psychologist*, **70**(2), 119. 1
- Weigelt, B., F. C. Geyer, and J. S. Reis-Filho** (2010). Histological types of breast cancer: how special are they? *Molecular oncology*, **4**(3), 192–208. 93
- Weyn, B., G. Van De Wouwer, A. Van Daele, P. Scheunders, D. Van Dyck, E. Van Marck, and W. Jacob** (1998). Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry: The Journal of the International Society for Analytical Cytology*, **33**(1), 32–40. 15

**WHO** (2020). Cancer key facts. <https://www.who.int/news-room/fact-sheets/detail/cancer>. 1

**Wolberg, W. H., W. N. Street, and O. L. Mangasarian** (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology*, **17**(2), 77–87. 15

**Wolff, A. C., M. E. H. Hammond, J. N. Schwartz, K. L. Hagerty, D. C. Allred, R. J. Cote, M. Dowsett, P. L. Fitzgibbons, W. M. Hanna, A. Langer, et al.** (2007). American society of clinical oncology/college of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Archives of pathology & laboratory medicine*, **131**(1), 18–43. 28

**Wollmann, T. and K. Rohr**, Deep residual hough voting for mitotic cell detection in histopathology images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017. 20

**Wu, J.** (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, **5**, 23. 19

**Xing, F., H. Su, J. Neltner, and L. Yang** (2013). Automatic ki-67 counting using robust cell detection and online dictionary learning. *IEEE Transactions on Biomedical Engineering*, **61**(3), 859–870. 29, 30

**Xu, J., C. Zhou, B. Lang, and Q. Liu**, Deep learning for histopathological image analysis: towards computerized diagnosis on cancers. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2017, 73–95. 23, 24, 84, 85, 87

**Yaziji, H., L. C. Goldstein, T. S. Barry, R. Werling, H. Hwang, G. K. Ellis, J. R. Gralow, R. B. Livingston, and A. M. Gown** (2004). Her-2 testing in breast cancer using parallel tissue-based methods. *Jama*, **291**(16), 1972–1977. 27

**Zaha, D. C.** (2014). Significance of immunohistochemistry in breast cancer. *World journal of clinical oncology*, **5**(3), 382. 5

**Zhao, X.-y., X. Wu, F.-f. Li, Y. Li, W.-h. Huang, K. Huang, X.-y. He, W. Fan, Z. Wu, M.-l. Chen, et al.** (2019). The application of deep learning in the risk grading of skin tumors for patients using clinical images. *Journal of medical systems*, **43**(8), 283. 19

**Zhong, F., R. Bi, B. Yu, F. Yang, W. Yang, and R. Shui** (2016). A comparison of visual assessment and automated digital image analysis of ki67 labeling index in breast cancer. *PloS one*, **11**(2), e0150505. 29

**Zoph, B., V. Vasudevan, J. Shlens, and Q. V. Le**, Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. 49

## LIST OF PAPERS BASED ON THESIS

### Journal Papers

1. **Tojo Mathew**, Jyoti R Kini, and Jeny Rajan, “Computational methods for automated mitosis detection in histopathology images: A review.” *Biocybernetics and Biomedical Engineering*, 41(1),64-82 (2021). [SCIE, IF: 4.31]
2. **Tojo Mathew**, B. Ajith, Jyoti R. Kini, and Jeny Rajan, “Deep learning based automated mitosis detection in histopathology images for breast cancer grading.” *International Journal of Imaging Systems and Technology*,1-17, (2022). [SCIE, IF: 2.00]
3. **Tojo Mathew**, S. Niyas, C. I. Johnpaul, Jyoti R. Kini, and Jeny Rajan, “A novel deep classifier framework for automated molecular subtyping of breast carcinoma using immunohistochemistry image analysis.” *Biomedical Signal Processing and Control*,76, 103657, (2022) [SCIE, IF: 3.88]
4. **Tojo Mathew**, B. Ajith, C. I. Johnpaul, Jyoti R Kini, and Jeny Rajan, “A deep learning based classifier framework for automated nuclear atypia scoring of breast carcinoma.” *The Breast*. (Under review) [SCIE, IF: 4.38]



# TOJO MATHEW

tojomathew@gmail.com

+91 9741599030

#16, Treasury Layout, Dattagalli-Bogadi, Mysuru, India, 570023

## EDUCATION

---

### National Institute of Technology Karnataka, Surathkal.

2016 - 2022

Doctor of Philosophy (Ph.D-Thesis submitted)

Coursework CGPA: 9.33/10

Department of Computer Science and Engineering

*Thesis Title:* Development of Deep Learning based Automated Methods for Breast Cancer Histopathology Image Analysis

*Research Advisor:* Dr. Jeny Rajan, Department of CSE, NITK Surathkal

### University of Kerala, Thiruvananthapuram, India

2012 - 2014

Master of Technology (M. Tech)

CGPA: 3.47/4

Department of Computer Science

*Specialization:* Digital Image Processing

*Thesis Title:* Reversible Data Hiding in Encrypted Images

### Mahatma Gandhi University, Kerala, India

1999 - 2003

Bachelor of Technology (B. Tech)

Percentage: 80.81

*Specialization:* Computer Science and Engineering

## WORK EXPERIENCE

---

### The National Institute of Engineering

Jan 2015 - Till date

*Assistant Professor*

*Mysuru, India*

### College of Engineering, Thiruvananthapuram

Jan 2014 - Jan 2015

*Guest Lecturer*

*Kerala, India*

### Teleca Software Solutions

June 2011 - Sept 2011

*Team Lead - Software development*

*Bengaluru, India*

### Nokia India

Feb 2008 - June 2011

*Senior Engineer - Software development*

*Bengaluru, India*

### Samsung India Software Operations

Apr 2005 - Nov 2007

*Senior Software Engineer*

*Bengaluru, India*

### Wipro Technologies

Oct 2003 - Mar 2005

*Project Engineer*

*Bengaluru, India*

## ACHIEVEMENTS

---

- ◇ Figured among top 2% in national-level NPTEL Online Certification Course (MOOC) **NBA Accreditation and Teaching and learning in Engineering (NATE)**, conducted by IISc Bangalore (Apr 2022) out of 967 candidates certified in this course.
- ◇ Figured among top 1% in national-level NPTEL Online Certification Course (MOOC) **C++ Programming**, conducted by IIT Kharagpur (Apr 2017) out of 2125 candidates certified in this course.
- ◇ **Best M.Tech thesis award** from Dept. of Computer Science, University of Kerala (Dec 2014).

- ◇ Qualified **National Eligibility Test by University Grants Commission (UGC-NET)** for Assistant Professor (Dec 2013)
- ◇ Qualified **Graduates Aptitude Test for Engineering (GATE)** in CS & Engg. with an All India Rank of 6163 out of 156780 candidates (Mar 2012).

## RESEARCH PUBLICATIONS

---

### SCIE Indexed Journals

- ◇ **Tojo Mathew**, Jyoti R. Kini, and Jeny Rajan. "Computational methods for automated mitosis detection in histopathology images: A review." *Biocybernetics and Biomedical Engineering (Elsevier)*, 2021, DOI: <https://doi.org/10.1016/j.bbe.2020.11.005> [SCIE, IF: 4.31].
- ◇ **Tojo Mathew**, Ajith B, Jyoti R Kini, and Jeny Rajan, "Deep Learning based Automated Mitosis Detection in Histopathology Images for Breast Cancer Grading", *International Journal of Imaging Systems and Technology (Wiley)*, 2022, DOI: <https://doi.org/10.1002/ima.22703> [SCIE, IF: 2.00].
- ◇ **Tojo Mathew**, C I Johnpaul, Jyoti R Kini, and Jeny Rajan, "An Ensemble Deep Classifier Framework for Automated Molecular Subtyping of Breast Carcinoma using Immunohistochemistry Image Analysis", *Biomedical Signal Processing and Control (Elsevier)*, 2022, DOI: <https://doi.org/10.1016/j.bspc.2022.103657> [SCIE, IF: 3.88].
- ◇ **Tojo Mathew**, C I Johnpaul, Ajith B, Jyoti R Kini, and Jeny Rajan, "A deep learning based classifier framework for automated nuclear atypia scoring of breast carcinoma", *The Breast (Under review)* [SCIE, IF: 4.38].

### International Conferences

- ◇ **Tojo Mathew** , and C I Johnpaul, "Reversible data hiding in encrypted images using interpolation-based distributed space reservation." In *Advanced Computing and Communication Systems (ICACCS)*, 2017 4th International Conference on, pp. 1-6. IEEE, (2017). DOI: <https://doi.org/10.1109/ICACCS.2017.8014608>
- ◇ C I Johnpaul, and **Tojo Mathew**, "NP-completeness of an optimization problem on plants selection using reduction method." *Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, International Conference on. IEEE, (2017). DOI: <https://doi.org/10.1109/ICICICT1.2017.8342556>
- ◇ C I Johnpaul, and **Tojo Mathew**, "A Cypher query based NoSQL data mining on protein datasets using Neo4j graph database." In *Advanced Computing and Communication Systems (ICACCS)*, 2017 4th International Conference on, pp. 1-6. IEEE, (2017) DOI: <https://doi.org/10.1109/ICACCS.2017.8014558>
- ◇ Nirmal S. Nair, **Tojo Mathew**, Neethu A. S., Viswajith P. Viswanath, Madhu S. Nair and Wilscy M., "A Proactive Approach to Reversible Data Hiding in Encrypted Images", *International Conference on Information and Communication Technologies - ICICT 2014 (CUSAT, Kochi, India)*, *Procedia Computer Science, Elsevier*, Vol.46, pp.1510-1517, (2015). DOI: <https://doi.org/10.1016/j.procs.2015.02.071>
- ◇ **Tojo Mathew**, Wilscy M. "Reversible Data Hiding in Encrypted Images by Active Block Exchange and Room Reservation", *International Conference on Contemporary Computing and Informatics - IC3I 2014 (Mysore, India)*, IEEE Computer Society Press, pp.839-844, (2014). DOI: <https://doi.org/10.1109/IC3I.2014.7019628>

## AREAS OF INTEREST

---

Medical image analysis, Deep learning, AI & ML, Digital image processing, Computer programming.

**Declaration:** I declare that the information stated above are true and valid to the best of my knowledge.

Tojo Mathew  
09 JUNE 2022