

CERTIFICATE

It is certified that the work contained in the thesis titled “*Artificial Intelligence Driven Techniques for Cancer Classification using Histopathological Images*” by *Abhinav* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.


Co-Supervisor
सहायक प्राध्यापक/Assistant Professor
समन्वय विज्ञान एवं अभियंत्रिकी विभाग/Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान (बनारस हिन्दू विश्वविद्यालय)
(बनारस हिन्दू विश्वविद्यालय)
Assistant Professor
बनारस /Varanasi-221005
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU)
Varanasi,
Uttar Pradesh, INDIA 221005.



Supervisor
पर्यवेक्षक/Supervisor
वैज्ञानिक विज्ञान एवं अभियंत्रिकी विभाग
Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान (बनारस हिन्दू विश्वविद्यालय)
Professor and HOD
Indian Institute of Technology (BHU)
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU)
Varanasi,
Uttar Pradesh, INDIA 221005.

DECLARATION BY THE CANDIDATE

I, **Abhinav**, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of **Prof. Sanjay Kumar Singh** from **July 2017** to **December 2021** at **Department of Computer Science and Engineering**, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.


Date:

Place: Varanasi


(Abhinav)

CERTIFICATE BY THE SUPERVISOR

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


(Prof. Sanjay Kumar Singh)
पदवेत्तक/Supervisor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Professor and HOD,
Dept. of Computer Science and Engineering
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(BHU) Varanasi
(Banaras Hindu University)
वाराणसी/Varanasi-221005

 21/12/21
Signature of Head of the Department
आचार्य व विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(वाराणसी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी-221005/Varanasi-221005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Artificial Intelligence Driven Techniques for Cancer Classification using Histopathological Images

Name of the Student: Abhinav

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

Date: 20.12.2021

Place: Varanasi


(Abhinav)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

This thesis is dedicated to my beloved family.

For their endless love, support and encouragement

ACKNOWLEDGEMENT

Apart from the efforts of ours, the success of any work depends largely on the encouragement and guidelines of many others. I take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this work. First of all, I would like to thank the Almighty for his unlimited favours, endless mercies, and choicest blessings bestowed upon me that helped me to come across all the way and imparadise.

At this stage, I owe my indebtedness to my supervisor, Prof. Sanjay Kumar Singh, Head of the Department, Computer Science and Engineering. He has been the torchbearer of my Ph.D. work. I express a deep sense of reverence to him for his invaluable guidance, constant encouragement, constructive suggestions, sagacious advice and keen supervision during the entire duration of my Ph.D. To have such an ingenuous perfectionist as my guide was a privilege for me as he gave new dimensions to my knowledge. A formal statement of acknowledgement will hardly meet the sense in the matter of expression of enormous debt I owe to him.

I further express my thankfulness to Dr Sonal and Dr Sameer Shrivastava, Department of Biotechnology, IVRI, for their prudent suggestions and for the kind consideration and cooperation which they extended whenever I approached them. A mere thanks and words of appreciation cannot compensate for their wholehearted support, incessant efforts, timely help, valuable suggestions, cordial encouragement, and generosity offered.

I would also like to take this opportunity to express my sincere regards to RPEC members Prof. Subir Das and Dr. Ravi Shankar Singh for their constant support and encouragement. I am also thankful to my co-supervisor Dr. Lakshmanan Kailasam, for the guidance and generosity offered. Regards are also due to all faculty members and staff of the Computer Science and Engineering Department for their kind support

and the necessary facilities provided. Further, I extend my special gratitude to my colleagues and friends, Dr. Anshul, Dr. Rishav Singh, Dr. Sandeep Sambhaji Udamle, Miss. Vandana, Mr. Ritesh Sharma, and Mr. Aneesh G Nath who supported me throughout my thesis work and provided technical, moral, and emotional support during my research work. I want to express my deepest gratitude to them for providing moral support throughout my thesis work. I have also spent some of the craziest and most memorable moments with them.

From the core of my heart, I thank my family for all that they have done in bringing me to this platform, and no words can describe my feeling towards them. My parents have been the driving force behind me, taking all pains to illumine my path and revive my energy. Their boundless love and encouragement are and always will be a constant source of motivation for me in my life. The constant encouragement, love and affection of my family have always been a reservoir of the impetus for me.


(Abhinav)

Contents

List of Figures	xi
List of Tables	xvi
List of Abbreviations	xvii
List of Symbols	xix
Preface	xxi
1 Introduction	1
1.1 Motivation	7
1.2 Challenges	10
1.3 Objectives	12
1.4 Thesis Contributions	13
1.5 Organization of the Thesis	15
2 Background and Related Works	17
2.1 Background	17
2.1.1 Breast cancer and CAD in cancer histopathology	18
2.1.2 Convolutional neural network	19
2.1.3 Transfer learning	22
2.2 Related Works	23
2.2.1 Feature learning	27
2.2.2 AI approaches for histopathological image classification	32
2.3 Summary	43
3 Data Acquisition and Processing	45
3.1 Introduction	45
3.2 CMT Tissue Processing and Image Acquisition Setup	47

3.3	Datasets Description	48
3.3.1	CMTHis dataset	48
3.3.2	BreakHis dataset	50
3.4	Data Pre-Processing Techniques	52
3.4.1	Colour and stain normalization	53
3.4.2	Data augmentation	56
3.4.3	Data splitting protocol	57
3.5	Summary	58
4	Deep Feature Learning using Convolutional Neural Network	59
4.1	Introduction	59
4.2	Motivation and Significant Contributions	60
4.3	Theoretical Background	62
4.3.1	Transfer learning	62
4.3.2	VGGNet-16 CNN	62
4.4	Proposed Methodology	65
4.4.1	VGGNet-16 as a feature extractor	65
4.4.2	Classifiers	67
4.5	Experimental Results and Analysis	68
4.5.1	Experimental setup	69
4.5.2	Performance evaluation of the proposed framework in magnification dependent model	71
4.5.3	Performance comparison with state-of-the-art CNN architectures	76
4.5.4	Effect of contemporary classifiers in magnification independent model	76
4.5.5	Effect of stain normalization	79
4.5.6	Ablation study	82
4.6	Summary	84
5	Secure Deep Feature Classification using Convolutional Neural Network	85
5.1	Introduction	85
5.2	Motivation and Significant Contribution	87
5.3	Theoretical Background	89
5.3.1	Privacy preserving in deep learning	89
5.3.2	Secure multiparty computation	90
5.3.3	Differential privacy	93

5.4	Proposed Methodology	95
5.4.1	VGGNet-16 as a feature extractor	95
5.4.2	Privacy preserved model training with differential privacy	97
5.4.3	Multi party secure tumor classification with cloud services	99
5.5	Experimental Results and Analysis	100
5.5.1	Experimental setup	100
5.5.2	Model evaluation	101
5.5.3	Comparison of the efficacy of the proposed framework with other CNN models	101
5.5.4	Application of differential privacy learning to the proposed model	105
5.5.5	Ablation study on Activation Function	107
5.6	Summary	108
6	Compression of Deep Convolutional Neural Network	111
6.1	Introduction	111
6.2	Motivation and Significant Contribution	112
6.3	Theoretical Background	114
6.3.1	Edge-based computing paradigms	114
6.3.2	Deep compression	115
6.3.3	MobileNet	117
6.4	Proposed Methodology	120
6.4.1	Model architecture	121
6.4.2	Model compression	123
6.5	Experimental Results and Analysis	124
6.5.1	Experimental setup	124
6.5.2	Evaluation metrics	125
6.5.3	Result analysis and discussion	126
6.5.4	Performance of the MobiHisNet on Raspberry Pi	136
6.6	Summary	137
7	Classification of Cancer Images using CoM-based Feature Descriptor	139
7.1	Introduction	139
7.2	Motivation and Significant Contribution	141
7.3	Theoretical Background	143
7.3.1	Basic definitions	143
7.3.2	Local patterns	144
7.3.3	Theoretical analysis of center of mass	146

7.3.4	l_1 – norm computation	148
7.3.5	Principal component analysis	149
7.3.6	Support vector machines	150
7.3.7	Fuzzy SVM	152
7.4	Proposed Methodology	155
7.4.1	Stain normalization	155
7.4.2	Feature extractor- CoMHisP	155
7.5	Experimental Results and Analysis	162
7.5.1	Experimental setup	162
7.5.2	Evaluation of the performance of CoMHisP in magnification dependent model	163
7.5.3	Effect of classifiers and stain normalization on the performance of the proposed framework in a magnification-independent model	173
7.6	Summary	174
8	Conclusion and Future Directions	177
8.1	Conclusion	177
8.2	Future Directions	181
	List of Publications	184
	Appendix	185
	Bibliography	189

List of Figures

1.1	Comparison between the workflow of traditional and digital pathology.	2
1.2	Generalize overview of the histopathological image classification.	6
2.1	A general overview of the CNN architecture with two convolution layers and one fully-connected layers.	20
2.2	An example of a convolution operation on a 7x7 matrix with a 3x3 kernel	20
2.3	An example of a max-pooling and an average-pooling operation. In this example, a 4x4 image is down-sampled to a 2x2 image by calculating the maximum or average value of each sub-region.	21
3.1	Image acquisition framework.	47
3.2	Magnifications of same tissue slide in the CMTHis dataset.	49
3.3	Representative H&E stained images of different classes present in CMTHis dataset.	50
3.4	Representative H&E stained images of different classes present in BreakHis dataset	51
3.5	Representative H&E stained histopathological images with and without stain normalization from the CMTHis dataset.	55
3.6	Sample images after applying data augmentation.	57
4.1	VGGNet-16 architecture	63
4.2	Proposed framework for histopathology image classification. H&E stained histopathological tumor images were initially subjected to data augmentation and stain normalisation. Finally, the images with and without stain normalisation were applied to the fused VGGNet-16 model for binary classification of cancer images, where all FC layers are removed.	66
4.3	Confusion matrix plots of BreakHis dataset at 200× magnification for (a) RF, (b) SVM(LIN), (c) SVM(RBF), and (d) SVM(POLY).	72

4.4	AUC for different classifiers at 200× magnification for (a) BreakHis dataset and (b) CMTHis dataset.	74
4.5	Mean test accuracy for different classifiers applied to proposed framework in magnification independent model.	78
4.6	Mean test accuracy for different classifiers of proposed framework with and without stain normalisation on BreakHis dataset.	80
4.7	Mean test accuracy for different classifiers of proposed frameworks with and without stain normalisation on CMTHis dataset.	81
4.8	Visualization of the feature maps of one of the images extracted from Block-1 to Block-5 in the VGGNet-16 Model.	83
5.1	Overall Framework	97
5.2	Representation of differentially private stochastic gradient descent (DPSGD) algorithm.	98
5.3	Representaion of private image classification framework.	100
5.4	Architecture of fully connected layers in the cloud.	102
5.5	The accuracy v/s epoch and loss v/s epoch curves of proposed model on the CMTHis dataset at 40X magnification.	103
5.6	Comparison of the accuracy of the proposed framework (w/o DP) on BreakHis and CMTHis datasets.	104
5.7	Effect of magnification on the efficacy of the proposed framework with and without DP.	106
5.8	Comparison of different activation functions on the efficacy of the proposed secure model on CMTHis Dataset.	108
6.1	Pruning	115
6.2	Quantization	116
6.3	Convolution operations	118
6.4	Illustration of depthwise separable convolution consisting of depthwise and pointwise convolution.	119
6.5	Proposed framework	120
6.6	Proposed model architecture	121
6.7	Effect of model compression on the accuracy, and model size.	129
6.8	Effect of model compression on inference time as well as on maximum memory used.	130
6.9	Effect of thread selection on inference time as well as on maximum memory used.	131

6.10	Comparison of on device inference and maximum memory utilization. .	135
6.11	(a) Raspberry Pi 3 B+ Rev 1.3 (1.4 GHz quad-core BCM2835 ARMv7 CPU with 1 GB RAM) Connected to Power Supply (b) Histopathological image testing process on Raspberry Pi.	137
7.1	Green dot represents center of mass of each shape.	146
7.2	l_1 -norm representation.	148
7.3	Overview of the proposed framework-CoMHisP.	156
7.4	Performance of the proposed CoMHisP feature extractor versus other state-of-the-art feature descriptors on the BreakHis dataset without stain normalization.	165
7.5	Bar diagram depicting classification accuracies achieved by applying the proposed CoMHisP feature extractor and different classifiers on CMTHis data after stain normalization at 40X magnification.	166
7.6	ROC curve analysis showing AUC of the CoMHisP framework using CMTHis dataset with different classifiers at 40X magnification.	167
7.7	Bar diagram depicting performance of the proposed CoMHisP framework with FSVM-WD classifier at different magnifications.	168
7.8	Performance of the proposed CoMHisP feature extractor versus other state-of-the-art feature descriptors on the CMTHis dataset without stain normalization.	169
7.9	Performance of the proposed CoMHisP feature extractor versus other state-of-the-art feature descriptors on the CMTHis dataset with stain normalization.	170
7.10	Effect of stain normalization on the performance of the proposed framework.	172

List of Tables

2.1	Overview of the intelligent analysis of breast cancer histopathology. . .	24
2.2	Description of object-level features used in CAD of histopathology images.	29
3.1	CMTHis dataset image distribution in terms of class and magnification factor.	49
3.2	BreakHis dataset image distribution in terms of class and magnification factor.	51
4.1	Mean test accuracy for different classifiers applied to proposed framework at different magnifications in the BreakHis dataset.	71
4.2	Mean test accuracy for different classifiers applied to proposed framework at different magnifications in the CMTHis dataset.	73
4.3	Comparison of our framework with state-of-the-art CNNs at different magnifications.	77
4.4	Comparison of the magnification independent model of proposed framework with other reported magnification independent models.	78
4.5	Ablation study on reducing the number of block in VGG16 and observing its effect on CMTHis dataset	82
5.1	Comparison of the model accuracy with other state-of-the-art methods.	105
6.1	Mobile devices used in the experiments	125
6.2	Comparative analysis of proposed MobiHisNet with state-of-the-art pretrained models.	128
6.3	Ablation study of proposed model with depth parameter γ	132
6.4	Ablation study of proposed MobiHisNet with state-of-the-art pretrained models	133
6.5	Ablation study of model comparison on device	134
6.6	Results of the proposed model with FP16 on Raspberry Pi	136

7.1	Performance of Different Feature Descriptors with Classifiers	164
7.2	Performance of Magnification-Independent Model using the CoMHisP Framework.	173

Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
AUC	Area Under Curve
CAD	Computer-Aided Diagnosis
CMT	Canine Mammary Tumour
CNN	Convolutional Neural Network
CoM	Center of Mass
DP	Differential Privacy
DPSGD	Differentially Private Stochastic Gradient Descent
FC	Fully Connected
GPU	Graphics Processing Unit
HBC	Human Breast Cancer
H&E	Hematoxylin and Eosin
IoT	Internet of Thing
LBP	Local Binary Patterns
MPC	Multi-party Computation
OD	Optical Density
RF	Random Forest
ROI	Region of Interest
SD	Standard Deviation

Abbreviation	Description
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SMPC	Secure Multi-party Computation

List of Symbols

Symbol	Description
ϱ	Angular feature vector
C_R	Compression rate
d_{jk}	Distance of j^{th} sample to the k^{th} neighbour
λ	Eigen value
K_i	Feature vector of i^{th} sample image
\vec{F}	Force acting on the system
g_t	Gradient
C_u	Gradient norm bound
G	Gradient of Center of Mass
g_x	Grey value of pixel
$L(i, j)$	Illuminance component
α	Lagrangian multiplier
γ_t	Learning rate
$L_1()$	l_1 -norm
m_i	Mass of i^{th} particle
μ	Mean value
Υ_i	Membership value of i^{th} sample
t_h	Number of corrupted participants
p	Number of neighbour pixel

Symbol	Description
σ	Noise scale
j	Objective function
C_{dws}	Overall operation cost for depth-wise convolutions
\vec{r}_{cm}	Position vector of Center of Mass
$h(\cdot)$	Pooling function
g_c	Reference gray value
$R(i, j)$	Reflectance component
$\rho()$	Relative density between two instances
C_{std}	Standard convolutional cost
s_f	Sensitivity
ξ_i	Slack variable for the i^{th} sample
T	Threshold value
D_i	Tissue images corresponding to i^{th} patient
γ	Width multiplier

Preface

Cancer is among the most perplexing ailments that has tormented humans for much of our civilization. In both humans and dogs, tumours of the mammary gland are most common among females. Both human breast cancer (HBC) and canine mammary tumours (CMTs) are associated with high mortality rates and poor prognosis. Moreover, the disease is similar in both the species and hence, CMTs are considered excellent models for HBC studies. Timely classification of cancers is crucial for deciding the most appropriate therapeutic regime and prognosis. However, many a times, there is a delay in diagnosis due to the paucity of trained oncopathologists. Accurate identification and classification of benign and malignant tumour tissues is a vital clinical task, and automated procedures can substantially speed up the diagnosis, minimize errors and contribute to better healthcare management. Thus, the development of automated computer-aided diagnosis (CAD) systems is a priority research area for image enhancement, feature extraction, and classification of cancer histopathological images.

The increasing adoption of deep learning and machine learning across healthcare domains, together with the availability of highly characterized cancer datasets, has accelerated research into the utility of artificial intelligence in the analysis of the complex biology of cancer. While early results are promising, this is a rapidly evolving field with new knowledge emerging in both cancer biology and deep learning. Deep learning-based approaches are recently being explored for analyzing histopathological images of HBC. However, so far, due to the lack of any publicly available CMT database, no studies

have focused on the automated classification of CMTs. The existence of heterogeneous and diverse types of CMTs and the paucity of skilled veterinary pathologists justify the need for automated diagnosis. Thus, the study was proposed with the aim to develop a framework for CAD of CMTs with high accuracy. To the best of our knowledge, for the first time, a dataset of CMT histopathological images (CMTHis) was introduced in this study. Further, a VGGNet-16 as a feature extractor was proposed, wherein a FC layer was removed and experimented with various classifiers.

As the success of deep learning depends on the amount of data available, the large-scale collection of data is critical for the effectiveness of the model. However, privacy and security concerns often prevent data owners from contributing sensitive information in machine learning systems that contain multiple parties. Thus, another objective of this study was to address such concerns related to confidential input of data during training or inference, as well as to the sharing of a trained model. To address these, an efficient and secure cancer diagnostic framework for histopathological image classification is proposed in this study by utilizing both differential privacy and secure multi-party computation.

Another area of concern addressed in this study is the poor applicability of deep learning-based approaches on edge devices. Moreover, deep learning-based approaches are computationally expensive and have huge parameters, which makes them less affordable for edge devices. Recently in the edge computing world, providing a low-cost solution has become highly valuable. To the best of our knowledge, this is the first study to develop a model for histopathological image classification that works effectively on edge devices. The proposed model was tested on a Raspberry Pi and three different smartphones to demonstrate its efficiency on a lightweight processor. The model proposed herewith speeds up histopathological image classification significantly without sacrificing accuracy, allowing it to compete with other models for applications on edge devices.

The requirement for massive amounts of labelled data is a major limiting factor for deep learning-based approaches. In the field of medical imaging, such large labelled datasets are difficult to acquire. Machine learning-based algorithms are a potential alternative to deep learning for medical imaging with smaller datasets. A major constraint of conventional machine learning techniques is the requirement of complex processing for the extraction of discriminatory features. Traditional machine learning algorithms have limited ability to reveal the most sophisticated features of cancer histopathological images, but their robustness and fault tolerance can be enhanced by using fuzzy modelling to capture the uncertainty in image data. Therefore, in this study, a novel CoMHisP framework based on a fuzzy support vector machine with within-class density information was also proposed. It utilizes a novel feature extraction technique by optimizing the block size to extract image micro-patterns and computing centre of mass (CoM) for each pixel to extract feature vectors. The proposed methodology is simpler and requires less processing time than other methods, and consequently does not require high-performance computers. It performs extraordinarily well at lower magnification, which is an additional advantage for providing diagnosis using simple microscopes in low-cost clinical settings.

The approaches, proposed in this study, have been evaluated on publicly available BreakHis dataset and newly introduced CMTHis dataset. The present study demonstrated effective solutions for the classification of HBC and CMTs with high accuracy and solves the concerns associated with deep learning-based approaches such as the requirement of large datasets, data security and privacy issues, as well as limited applicability on edge devices.