

**HISTOPATHOLOGICAL BREAST-CANCER IMAGE CLASSIFICATION
BASED ON MACHINE-LEARNING TECHNIQUES**

by

Abdullah-Al Nahid



Dissertation submitted in partial fulfilment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering
Faculty of Science and Engineering
Macquarie University
Sydney, Australia

April 2018

STATEMENT OF CANDIDATE

I certify that the work in this thesis entitled "Histopathological Breast-Cancer Image Classification Based on Machine-Learning Techniques " has not previously been submitted for a degree, nor has it been submitted as part of the requirements for a degree to any university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me.

In addition, I certify that all information sources and literature used are indicated in the thesis.

.....

Abdullah-Al Nahid

To my adorable son

Arham Juhas Rihun

ACKNOWLEDGMENTS

This thesis would not have been possible without the help, support and patience provided by my principal supervisor Dr Yinan Kong in supervising my PhD candidature. My PhD work is funded by an International Macquarie University Research Excellence Scholarship (iMQRES), Macquarie University, Australia. A very special thanks to Macquarie University for providing me with this prestigious scholarship. I would like to thank my fellow labmates for their help. I am grateful to Dr Keith Imrie for his wonderful support and patience in proof-reading my articles and thesis. I would like to thank Dr Leonard G.C. Hamey, senior lecturer in the Department of Computing at Macquarie University for introducing me to the state-of-the-art Deep-Learning techniques.

I would like to thank my family members for their support. I started my PhD in April, 2015, and my son Arham Juhas Rihun was born in December 2015. Throughout my PhD journey my son lived in Bangladesh with his mother. For more than two years I have missed my son, and thank him for his sacrifice.

ABSTRACT

Machine-Learning (ML) techniques bring a new paradigm which has generated a revolutionary momentum and wrought changes in every day of our modern lives, ranging from autonomous lifestyles to decision-making scenarios. Among the different branches of ML activities, the medical field is notable, covering the field of detection and monitoring as well as the present status of diseases. Among the different medical diseases cancer is a serious threat. In particular, breast cancer is always a serious threat to women. Proper identification and then proper management and monitoring help the patient to recover from the disease, or at least help them to lead a better life. Proper identification and the current status of cancer largely depend on biomedical image analysis, a complex area of understanding. The analysis of these images requires special knowledge. The autonomous finding of Benign and Malignant information based on the images and making a Computer-Aided Diagnosis (CAD) system provide both the patient and the doctor with a second layer of confidence and allow them to make a more reliable decision. For the autonomous identification and detection of cancer, digital ML techniques have provided a revolutionary improvement. The recent development of the Deep Neural Network (DNN) and the logic-based algorithm make it possible to detect the target form the image more reliably. In this thesis we have investigated the performance of the DNN-based biomedical image classifier as well as the Extreme Gradient Boosting (XGBoost)-based image classifier for the autonomous CAD system.

Contents

Table of Contents	xi
List of Figures	xvii
List of Tables	xxv
List of Publications	xxix
1 Introduction	1
1.1 Motivation	3
1.2 Methodology	5
1.3 Thesis Outline	8
1.4 Chapter-wise Contributions	10
1.5 Author's Contributions	14
2 Involvement of Machine Learning for Breast-Cancer Image Classification: A Survey	17
2.1 Abstract	17
2.2 Introduction	18
2.3 Breast-Image Classification	22
2.3.1 Available Breast-Image Databases	23

2.3.2	Feature-Extraction and Selection	26
2.3.3	Classifier Model	30
2.3.4	Performance-Measuring Parameters	31
2.4	Performance of different Classifier models on Breast-Image Dataset	32
2.4.1	Performance Based on Supervised Learning	32
2.4.2	Performance Based on Un-supervised Learning	66
2.4.3	Performance Based on Semi-supervised Learning	70
2.5	Conclusion	73
3	Histopathological Breast-Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering	77
3.1	Abstract	77
3.2	Introduction	78
3.3	Feature partitioning	82
3.4	Deep Neural Network	84
3.4.1	Convolutional Neural Network	85
3.4.2	LSTM	90
3.4.3	CNN-LSTM	92
3.5	Proposed Models	93
3.5.1	Model-1	93
3.5.2	Model-2	95
3.5.3	Model-3	95
3.6	Results and Discussion	96
3.6.1	Performance of different Models	96
3.6.2	Effect of TS and ID	108
3.6.3	The effect of Cluster size (K) and Bandwidth (BW)	109
3.7	Recent findings for Breast-Image Classification based on DNN	111

3.8	Conclusion	113
4	Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network	117
4.1	Abstract	117
4.2	Introduction	118
4.3	Overall Architecture for Classification	121
4.3.1	Retinex Algorithm	122
4.4	Utilised Convolutional Neural Network	123
4.5	Results and Discussion	129
4.5.1	Time and Parameters required	133
4.6	Conclusion	135
5	Local and Global Feature Utilisation for Breast-Image Classification by Convolutional Neural Network	137
5.1	Abstract	137
5.2	Introduction	138
5.3	Overall Architecture for Classification	140
5.3.1	Handcrafted Features	141
5.4	Convolutional Neural Network and our proposed model	142
5.4.1	Our Model for Classification	146
5.5	Results and Discussion	146
5.5.1	Performance-Measuring Parameters	146
5.5.2	Results and Discussion	147
5.6	Conclusion	152
6	Frequency-Domain Information along with LSTM and GRU Methods for Histopathological Breast-Image Classification	153

6.1	Abstract	153
6.2	Introduction	154
6.3	Overall Architecture	156
6.4	LSTM and GRU Methods	159
6.4.1	Our Model	161
6.5	Results and Discussion	162
6.6	Conclusion	169
7	Histopathological Breast-Image Classification using Local and Frequency domains by Convolutional Neural Network	171
7.1	Abstract	171
7.2	Introduction	173
7.3	Overall Architecture	178
7.4	Feature Extraction and Data Preparation	181
7.4.1	Data Preparation for Case2	181
7.4.2	Data Preparation for Case3	184
7.5	Convolutional Neural Network	185
7.5.1	CNN Model for Image Classification	189
7.6	Performance-Measuring Parameters and utilised Platform	191
7.7	Results and Discussion	193
7.7.1	Performance of 40× Dataset	193
7.7.2	Performance of 100× Dataset	196
7.7.3	Performance of 200× Dataset	198
7.7.4	Performance of 400× Dataset	200
7.7.5	Required Time and Parameters	203
7.7.6	Comparison with Findings	203
7.8	Conclusion	206

8	Histopathological Breast-Image Classification with Restricted Boltzmann Machine along with Back Propagation	207
8.1	Abstract	207
8.2	Introduction	208
8.3	Image-Classification Model	212
8.4	Proposed RBM Model for Image Classification	213
8.5	Contrast-Enhancement	217
8.6	Feature-Extraction	220
8.7	Results and Discussion	223
8.7.1	Results and Comparison	224
8.8	Conclusion	234
9	Histopathological Breast-Cancer Image Classification with Feature Prioritisation	235
9.1	Abstract	235
9.2	Introduction	236
9.3	Classification Methodology	240
9.3.1	Extreme Gradient Boosting	241
9.3.2	Feature-Extraction	242
9.4	Comparison and Explanation of XGBoost	247
9.4.1	Performance based on XGBoost algorithm along with Tamura Features	255
9.5	Feature-Selection Methodology	257
9.5.1	Filter	258
9.5.2	Wrapper	266
9.6	Conclusion	273

10 Conclusion and Future Work	275
10.1 Conclusion	275
10.2 Future Research Directions	278
11 List of Abbreviations	281
References	285

List of Figures

1.1	Chapter-wise organisation	9
2.1	Numbers of new people facing cancer in Australia from 2007 to 2018 [1]. .	19
2.2	Numbers of people dying due to cancer in Australia from 2007 to 2018 [1].	19
2.3	Anatomy of the female breast images (For the National Cancer Institute © 2011; Terese Winslow, U.S. Govt, has certain rights)	20
2.4	(a,b) show mammograms of benign and malignant images (Examples of non-invasive image) and (c,d) show histopathological benign and malignant images (Examples of invasive image)	22
2.5	A very basic Breast-Image Classification Model	23
2.6	Number of papers published based on MIAS and DDSM databases	25
2.7	Classification of Features for Breast-Image Classification	26
2.8	A summary of Feature-Selection Method	31
2.9	Confusion Matrix	31
2.10	A generalised Supervised Classifier Model	33
2.11	A model of a Biological neuron	34
2.12	Working principle of a simple Neural Network technique	35
2.13	ReLU Operation	41
2.14	Max-Pooling and Average Pooling	42

2.15	Workflow of a Convolutional Neural Network	43
2.16	A general structure of a Tree	51
2.17	SVM finds the hyperplane which separates two classes	55
3.1	Overall image-classifier model for benign and malignant image classification	81
3.2	Figures a, b, c represent an original benign image, the KM cluster-transformed image, and the MS cluster-transformed image, respectively. Figures d, e, f represent an original malignant image, the KM cluster-transformed image and the MS cluster-transformed image, respectively.	84
3.3	Sigmoid, TanH, ReLU and Leaky ReLU	86
3.4	Pooling operation performed by 2×2 kernel	87
3.5	Drop-out	88
3.6	Workflow of a Convolutional Neural Network	89
3.7	A generalised RNN model, where the RNN output is computed and the reference information passes through the hidden unit	90
3.8	A generalised cell structure of an LSTM	91
3.9	CNN and LSTM models combined	92
3.10	Conventional CNN, LSTM-based architecture (a,b), CNN-LSTM-based architecture (c)	94
3.11	Statistical breakdown of the BreakHis dataset.	97
3.12	Comparison of Accuracy between Model-1, Model-2 and Model-3	101
3.13	Comparison of Precision between Model-1, Model-2 and Model-3	103
3.14	Comparison of F-Measure between Model-1, Model-2 and Model-3	104
3.15	Accuracy, Loss and M.C.C. values for Model-1 when we utilised the $40 \times$ dataset MS and Softmax together	106
3.16	Accuracy, loss and M.C.C. values for Model-2 when we utilised the $200 \times$ dataset, MS and Softmax together	107

3.17 Accuracy, loss and M.C.C. values for Model-3 with the $200\times$ dataset, KM and Softmax together	108
3.18 Figures a, b, c and d represent the Accuracy of the $40\times$, $100\times$, $200\times$, $400\times$ datasets for Model-2 with varying TS and ID	110
4.1 Number of Female deaths in Australia since 2012 due to breast cancer . . .	118
4.2 Overall image-classification model	121
4.3 Workflow of a Convolutional Neural Network	123
4.4 Conventional Model	125
4.5 CNN model with Residual Block	126
4.6 Max-Min Model	128
4.7 (a), (b), (c) and (d) represent the training and test accuracy comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.	131
4.8 (a), (b), (c) and (d) represent the training and test loss comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.	133
4.9 (a), (b), (c) and (d) represent the training and test M.C.C comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.	134
5.1 Male and Female cancer death statistics for the last decade in Australia. .	138
5.2 Overall image-classification model	141
5.3 Workflow of a Convolutional Neural Network	143
5.4 ReLU Opearation	144
5.5 The CNN model utilised for the classification	145
5.6 Confusion Matrix	147
5.7 Accuracy information when we utilise the C-H algorithm on the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively	149

5.8	M.C.C. information when we utilise the C-H algorithm on the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively	151
6.1	Female deaths from breast cancer: statistics for the last decade in Australia.	154
6.2	Overall Architecture for Breast-image classification	157
6.3	Feature preparation for LSTM/GRU model when FFT is utilised	158
6.4	Feature preparation for LSTM when DCT is utilised	159
6.5	Conventional NN and RNN models	160
6.6	A generalised cell structure of an LSTM	161
6.7	The proposed model where the cells (LSTM/GRU) are stacked	162
6.8	Number of parameters and required time for individual cases	166
6.9	(a), (b), (c) and (d) show the Accuracy, Loss, Kulback Divergence and Matthews correlation coefficient values for Case-5 on the $200\times$ dataset . .	167
6.10	(a), (b), (c) and (d) show the Accuracy, Loss, Kulback Divergence and Matthews correlation coefficient values for Case-8 on the $200\times$ dataset . .	168
7.1	New cases of breast cancer for women and number of women dying in the last twelve years	173
7.2	Left side represents Benign and right side Malignant histopathological images (This data has been collected from the BreakHis dataset)	179
7.3	Overall image-classification model.	179
7.4	(a) Wedge-shaped frequency response for 4-band decomposition and (b) Contourlet Transform working mechanism	182
7.5	Feature-Selection Procedure from images when we use DFT DCT	185
7.6	This figure represents the effects of kernel size, the size of stride and zero padding in a convolutional operation	187
7.7	Workflow of a Convolutional Neural Network	188

7.8	Architecture of Model-1 at the left and architecture of Model-2 at the right	191
7.9	Confusion Matrix	192
7.10	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the 40× dataset.	195
7.11	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the 40× dataset.	195
7.12	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the 100× dataset.	197
7.13	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the 100× dataset.	198
7.14	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the 200× dataset.	199
7.15	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the 200× dataset.	200
7.16	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the 400× dataset.	202
7.17	(a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH case on the 400× dataset.	202
8.1	Death statistics due to BC for the last 5 years in Australia	209
8.2	Workflow of Algorithm-1	212
8.3	Workflow of Algorithm-2	213
8.4	Graphical representation of BM and RBM models	214
8.5	DBN model for analysis of the data	216
8.6	Block Diagram of Step-1.	217
8.7	Block Diagram of Step-2.	218
8.8	Block Diagram of Step-3.	218

8.9	Tamura features extraction from the three different channels	220
8.10	(a), (b), (c) and (d) represent the Confusion Matrices for Algorithm-1 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.	225
8.11	(a), (b), (c) and (d) represent the Confusion Matrices for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.	226
8.12	(a), (b), (c) and (d) represent the performance analysis for Algorithm-1 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively. . .	229
8.13	(a), (b), (c) and (d) represent the performance analysis for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively. . .	230
8.14	(a), (b), (c) and (d) represent the ROC curves for Algorithm-1 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.	232
8.15	(a), (b), (c) and (d) represent the ROC curves for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.	233
9.1	Death statistics for Breast Cancer over the last decade in Australia	237
9.2	(a), (b) show Benign and Malignant images selected from the BreakHis dataset.	241
9.3	A very basic image-classification model	241
9.4	Overall Feature-Extraction from the three different channels	243
9.5	(a), (b), (c) and (d) show the ROC curves for the LBP, Tamura, Histogram and Harlick features, respectively, when we utilise the XGBoost algorithm.	254
9.6	(a) shows the Log Loss performance against the Number of Trees along with the depth of tree. (b) shows the Log Loss value for different learning rates when the depth of tree is 5 and the Number of Trees is 1050.	256
9.7	(a) shows the ROC curve and (b) represents the Confusion Matrix when the depth of tree is 5, Number of Trees is 1050, and learning rate is 0.10	257
9.8	A wrapper method for Feature-Selection	266

9.9	(a), (b), (c), (d), (e) and (f) show the Specificity, Recall, Precision, F-measure, Accuracy and model construction time for different feature sets based on different Feature-Selection algorithms.	272
-----	--	-----

List of Tables

1.1	Deaths due to cancer worldwide in 2015 [WHO]	3
1.2	Individual Contributions	15
2.1	Available Breast-Image Database for Biomedical Investigation	24
2.2	Feature Descriptor (Part 1)	28
2.3	Feature Descriptor (Part 2)	29
2.4	A simplified Hierarchy of Classification	36
2.5	Neural Network for Breast-Image Classification (Part 1)	37
2.6	Neural Network for Breast-Image Classification (Part 2)	38
2.7	Neural Network for Breast-Image Classification (Part 3)	39
2.8	Available Software for Deep-Learning Analysis	44
2.9	Convolutional Neural Network (Part 1)	47
2.10	Convolutional Neural Network (Part 2)	48
2.11	Convolutional Neural Network (Part 3)	49
2.12	Logic-Based (Part 1)	53
2.13	Logic-Based (Part 2)	54
2.14	SVM for Breast-Image Classification (Part 1)	59
2.15	SVM for Breast-Image Classification (Part 2)	60
2.16	SVM for Breast-Image Classification (Part 3)	61

2.17 Bayesian Classifier (Part 1)	64
2.18 Bayesian Classifier (Part 2)	65
2.19 K-means Cluster Algorithm and Self-Organising Map for Breast-Image Classification (Part 1)	68
2.20 K-means Cluster Algorithm and Self-Organising Map for Breast-Image Classification (Part 2)	69
2.21 Semi-Supervised Algorithm for Breast-Image Classification (Part 1)	71
2.22 Semi-Supervised Algorithm for Breast-Image Classification (Part 2)	72
3.1 Cancer Statistics for Australia 2017 [2]	78
3.2 Comparison of TN, FP, FN and TP values % for the different algorithms and different datasets	98
3.3 Average time and Parameters for various TS and ID	108
3.4 Effect of the cluster size (K) and the Bandwidth (BW)(%)	111
3.5 CNN and Histopathological findings	112
3.6 Comparing Accuracy (%) in different models	113
4.1 Performances of the different Cases on different datasets	129
4.2 Time (per/epoch) and parameters required to run the models	135
5.1 The performance of the C-H and C-L algorithms on the BreakHis dataset .	148
6.1 Description of the Cases	163
6.2 Performances of the different Cases on different datasets	164
7.1 Number of Handcrafted Features	186
7.2 A summary of classification-performance measurement parameters	192
7.3 Performance of various cases on 40× dataset	194
7.4 Performance of various cases on 100× dataset	196

7.5	Performance of various cases on $200 \times$ dataset	199
7.6	Performance of various cases on $400 \times$ dataset	201
7.7	Required Time and Number of Parameters	203
7.8	Summarises a few recent findings of Histopathological breast-image classification	205
8.1	Detailed description of each block of the machine	216
8.2	Few Performance measuring parameters along with CM	224
8.3	MSE values and the corresponding epoch values	228
8.4	Comparison of results using our proposed algorithm and other algorithms .	234
9.1	Harlick Features mathematical representation	246
9.2	Initial Parameters for the XGBoost Algorithm	248
9.3	Comparison of various classifiers with XGBoost algorithm using Tamura features	249
9.4	Comparison of various classifiers with XGBoost algorithm using Histogram features	250
9.5	Comparison of various classifiers with XGBoost algorithm using LBP features	251
9.6	Comparison of various classifiers with XGBoost algorithm using Harlick features	252
9.7	Overall best classifiers	253
9.8	Accuracy and Model Construction Time	255
9.9	Average Recall/Accuracy/Precision/Specificity when depth of tree is 5 and Number of Trees is 1050	257
9.10	Feature Priority table for various Feature-Selection Methods	260
9.11	All the feature sets based on the Chi-Square (CS) Feature-Selection method	262
9.12	All the feature sets based on the Fisher Score (FS) Feature-Selection method	263

9.13 All the feature sets based on the Relief (RE) Feature-Selection method . .	264
9.14 All the feature sets based on the Mutual Information (MI) Feature-Selection method	265
9.15 All the feature sets based on the Forward Feature-Selection method	268
9.16 All the feature sets based on the Backward Feature-Selection method . . .	269

List of Publications

Published journal articles included in this thesis:

- A. A. Nahid, Y. Kong, “Involvement of Machine Learning for Breast Cancer Image Classification: A Survey”, *Computational and Mathematical Methods in Medicine, Hindawi*, pp. 1–29, vol. 2017. [Online].
Available: <https://doi.org/10.1155/2017/3781951>.
- A. A. Nahid, Y. Kong, “Histopathological Breast-Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering”, *BioMed Research International, Hindawi*, pp. 1-20, 2018. [Online].
Available: <https://www.hindawi.com/journals/bmri/2018/2362108/>.
- A. A. Nahid, Y. Kong, “Histopathological Breast-Image Classification Using Local and Frequency Domains by Convolutional Neural Network”, *Information, MDPI*, vol. 9, no. 1, pp. 1–26, 2018. [Online].
Available: <http://www.mdpi.com/2078-2489/9/1/19>.
- A. A. Nahid, A. Mikaelian and Y. Kong, “Histopathological Breast Image Classification with Restricted Boltzmann Machine along with Back Propagation”, *BioMed Research, Allied Academics*.
Available: [Accepted on 2nd April, 2018].

Submitted journal articles which are included in this thesis:

- A. A. Nahid, A. Mikaelian, M. A. Mehrabi and Y. Kong, “Histopathological Breast-Cancer Image Classification with Feature Prioritisation”, *BioMed Research International, Hindawi*. [in review].

Published journal articles not included in this thesis:

- A. A. Nahid, T. M. Khan, and Y. Kong, “Hardware Implementation of Bone Fracture Detector Using Fuzzy Method Along with Local Normalization Technique”, *Annals of Data Science, Springer*, pp. 533–546, July 2017. [Online]. Available: <https://link.springer.com/article/10.1007/s40745-017-0118-z>.

Submitted journal articles which are not included in this thesis:

- A. A. Nahid, and Y. Kong, “Histopathological Breast Image Classification using Concatenated R-G-B Histogram Information”, *Annals of Data Science, Springer*. [in review].
- A. A. Nahid, and Y. Kong, “Noise Reduction in Biomedical Breast Images Using Adaptive Bilateral Filter”, *BioMed Research, Allied Academics*. [in review].

Published conference papers included in this thesis:

- A. A. Nahid, F. B. Ali, and Y. Kong, “Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network”, in *2017 20th International Conference on Computer and Information Technology (ICCIT), IEEE*, pp. 1-6, Dec. 2017. Available: <http://ieeexplore.ieee.org/document/8281815/>.
- A. A. Nahid and Y. Kong, “Local and Global Feature Utilization for Breast Image Classification by Convolutional Neural Network”, in *2017 International Conference*

on *Digital Image Computing: Techniques and Applications (DICTA)*, *IEEE*, pp. 1-6, Nov. 2017

Available: <http://ieeexplore.ieee.org/document/8227460/>.

- A. A. Nahid, M. A. Mehrabi and Y. Kong, “Frequency-Domain Information Along with LSTM and GRU Methods for Histopathological Breast-Image Classification”, in *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, *IEEE*, pp. 1-6, Dec. 2017.

Available: —————.

Published conference papers not included in this thesis:

- A. A. Nahid, T. M. Khan, and Y. Kong, “Performance Analysis of Integrated Canny and Fuzzy-Logic Based (2-by-2 Cell Block) Edge-Detection Algorithms”, in *2016 European Modelling Symposium (EMS)*, *IEEE*, pp. 64-69, Nov. 2016.
Available: <http://ieeexplore.ieee.org/abstract/document/7920230/>.
- A. A. Nahid, T. M. Khan, and Y. Kong, “Breast Image Classification Based on Concatenated Statistical, Structural and Textural Features”, in *2016 European Modelling Symposium (EMS)*, *IEEE*, pp. 27-32, Nov. 2017.
Available: <http://ieeexplore.ieee.org/abstract/document/7920224/>.
- A. A. Nahid, Y. Kong, and M. S. Hossain, “Performance Analysis of Integrated Canny and Fuzzy Logic Based (3-by-3 Cell Block) Edge Detection Algorithms”, in *2015 IEEE International Conference on Data Science and Data Intensive Systems, IEEE*, pp. 190-195, Dec. 2017.
Available: <http://ieeexplore.ieee.org/abstract/document/7396502/>.
- A. A. Nahid, Y. Kong, and M. N. Hasan, “Performance Analysis of Canny’s Edge Detection Method for Modified Threshold Algorithms”, in *2015 International Con-*

ference on Electrical & Electronic Engineering (ICEEE), IEEE, pp. 93-96, Nov. 2017.

Available: <http://ieeexplore.ieee.org/abstract/document/7428227/>.

- A. A. Nahid, Y. Kong, and M. N. Hasan, "Performance Analysis of Canny Methods for Various Thresholds", *in 2015 IEEE Student Conference on Science & Engineering (SCSE), University of Dhaka, Bangladesh, Sep. 2015.*
- A. A. Nahid, Y. Kong, and M. N. Hasan, "A Short Review on Edge Detection Techniques", *in IEEE Student Conference on Science & Engineering (SCSE), University of Dhaka, Bangladesh, Sep. 2015.*

Chapter 1

Introduction

Learning through observation or by instruction is a natural phenomenon which is common in wildlife and human beings. People try to learn from observation, interpret this observation and then apply this knowledge for better decisions in their day-to-day activities. Machine-Learning (ML) mimics this concept with the help of an intelligent machine (computer), providing a decision or suggestion on how to react to a new situation. In general the term intelligence is more related to human intellectual thinking ability. This concept about intelligence raises the following question: "*can a machine think and provide intelligent output?*". The answer is that a machine can think in an intelligent way if a proper mathematical and statistical knowledge of human thinking for a particular understanding is transferred to the machine through proper programming, such that the machine can mimic human thinking ability. This can allow a machine to learn and provide a decision. Basically a contribution of the computer sciences, mathematical and statistical approaches allow ML to assist a human to relate the present with the past, and wields an influence on almost all aspects of modern life.

The pioneer of ML, Arthur Lee Samuel, defined it as "*—field of study that gives computers the ability to learn without being explicitly programmed*". The recent best applica-

tions of ML are autonomous car driving, Google Glass, tagging photos using Facebook, voice synthesiser, home assistance like Google Home, and many more applications.

As the ML field has grown rapidly, with some advanced mathematical structures and model utilisation, ML algorithms have been adjusted to analyse and classify biomedical data. Gathering knowledge from biomedical big data, discovering information from the data, and adjusting that information for decision-making purposes, is of great interest. This includes biomedical images for disease analysis, identification and classification, which have generated a high level of interest in the research community.

Among various diseases, Breast-Cancer (BC) poses a serious threat to women. Diagnosis of BC largely depends on investigating biomedical images such as X-Ray and Histopathological images. Like a few other images, Histopathological images and biomedical images are of huge importance to experts, as they reveal detailed characteristics of the disease, and also deliver valuable information about the current status of the BC. However, Histopathological images are very complex in nature and need subjective knowledge to find meaningful information from them. The lack of specialists working in this field increases patients' waiting time, and sometimes physicians might disagree with others about the diagnosis. However, state-of-the-art ML has a great potential for biomedical image analysis, especially feature detection and classification. Considering the devastation wrought by BC, this thesis investigates and classifies a set of Histopathological images into Benign and Malignant classes. As a tool, ML techniques, specifically the state-of-the-art Deep Neural Network (DNN) and Extreme Gradient Boosting Algorithm (XGBoost) have been utilised for Histopathological BC image classification.

1.1 Motivation

The cell is considered to be the basic working unit of life. It is a natural phenomenon that cells of the body will generate and die to maintain the primitive working instrument; this creation of cells is normally controlled within a very tight constraint. However, sometimes this constraint is violated and situations occur when cells behave abnormally. According to the United States Cancer National Institute, *"Cancer is a disease in which abnormal cells divide without control and are able to invade other tissues"*. In addition to this, cancer is not considered as a single disease, rather it is a collection of different diseases. As stated in *www.cancer.org*, *"Cancer (CAN-sur): a word used to describe more than 100 diseases in which cells grow out of control; or a tumor with cancer in it"*.

Recent statistics show that deaths due to cancer were exceeded only by those of cardiovascular diseases throughout the world. Among the different cancer diseases Lung, Liver, Colorectal, Stomach and Breast cancer are the leading ones; statistics for the year 2015 are presented in Table 1.1, which shows that deaths due to BC were in fifth place among the prominent cancer diseases. Both men and women are vulnerable to BC. However, the probability of BC in women is higher than in men due to how women are structured physically.

Table 1.1: Deaths due to cancer worldwide in 2015 [WHO]

Cancer Type	Number of Deaths
Lung	1690000
Liver	788000
Colorectal	774000
Stomach	754000
Breast	571000

Cancer-affected women lead a life which may end in an early death. However, early detection as well as monitoring of the current status of the BC can improve the present BC scenario. The identification of the cancer largely depends on the analysis of biomed-

ical photographs. Modern medical photographic techniques such as X-Ray, Ultrasound, and Magnetic Resonance Imaging (MRI) have been used to take a photograph from the targeted area. These kinds of photographic techniques provide preliminary information about the BC, particularly the lymph cells. However, this preliminary information is not sufficient to provide a meaningful and detailed diagnosis of the BC. To get more reliable and accurate information, sample tissues from the affected area are collected and investigated, which is known as biopsy. The microscopic images of the biopsy tissues are captured for further investigation, and are known as Histopathological images. Though this kind of imaging technique is invasive, the images do provide a significant amount of information about the current status of the cancer.

Primarily, BC physicians, surgeons or doctors investigate the Histopathological images and provide a diagnosis. Gathering accurate knowledge from the biomedical data is a challenging task because it demands special knowledge and skills which are scarce. Sometimes the process of getting opinions about the disease based on the information from biomedical Histopathological images are so complex that specialists might disagree. Due to the above-mentioned issues, Computer-Aided Diagnosis (CAD) techniques help the doctor to offer a reliable opinion about the cancer. Since a decision provided by the CAD system is directly relevant to human life, the system needs to be designed with the utmost care. As time goes on, new methodologies and techniques are introduced which provide more reliable and accurate decisions.

1.2 Methodology

This thesis will discuss the design of an autonomous CAD system which classifies a set of Histopathological images into Benign and Malignant classes with ML techniques. The research has been carried out employing the following steps:

- Data Collection:

Image classification is performed on a dataset or a set of datasets. All the experiments in this thesis have been conducted on the BreakHis BC image dataset, which has been contributed by Professor Luiz Eduardo S. Oliveria, Federal University of Parana, Department of Informatics. This dataset contains 7909 Histopathological images which are further subdivided into four clusters $\{40\times, 100\times, 200\times, 400\times\}$ depending on the magnification factor.

- Feature-Extraction:

Features of a particular object allow the machine to learn characteristics about the data, and later, based on these characteristics, a classifier model is created to classify the data. The conventional method of Feature-Extraction is to locally handcraft the individual properties based on some criteria. This technique requires feature-specific knowledge to extract the appropriate information, and mostly these features belong to a "shallow" model. Another way of feature learning is learning about the feature in a hierarchical way, where the model learns the features from a low level to a complex level, which is known as a global feature. Global features are mostly utilised for the Convolutional Neural Network (CNN)-based model, where kernel methods serve to extract the information. In this thesis, we have used both locally extracted handcrafted features as well as global feature information. For the handcrafted local Feature-Extraction we have utilised:

1. Local Binary Pattern (LBP, Visual Descriptor): This provides information

about textural feature distribution.

2. Histogram: A statistical description of the pixels in an image is presented by a Histogram, which provides significant information about the image properties.
3. Tamura: Low-level statistical properties such as Tamura features have been extracted and utilised for image classification.
4. Harlick: Harlick features utilise a Gray-Level Co-Occurrence Matrix (GLCM) for calculating the textural features.

This research also utilises a few transformation-based methods to extract the properties from the images such as:

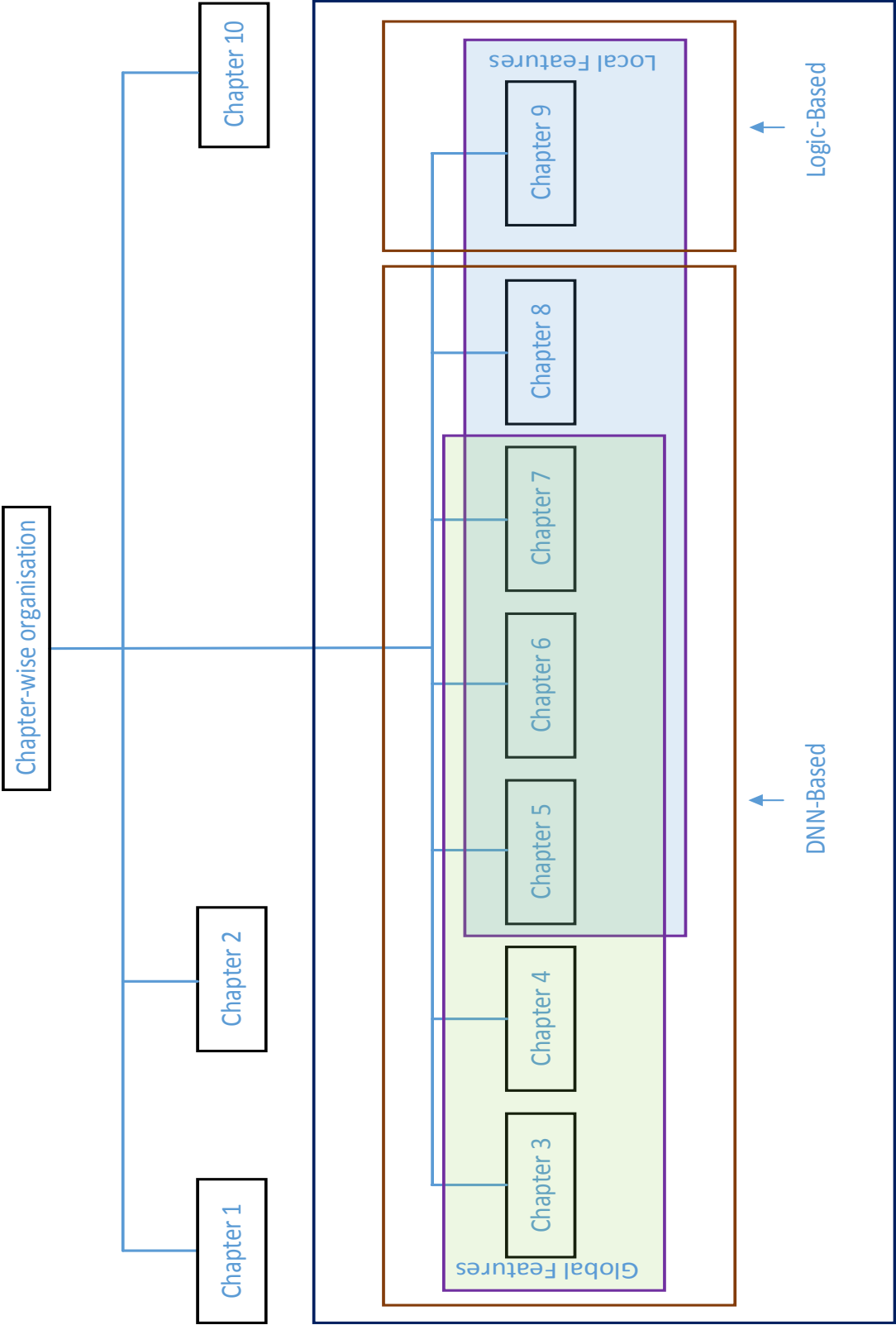
1. Contourlet Transform (CT) : Due to its mathematical structure, the conventional Wavelet Transform (WT) ignores the smoothness properties around the contour. To include information about the smoothness properties we have utilised CT.
 2. Frequency-domain information: Frequency-domain information has been extracted utilising Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) techniques.
- Classifier Model: Data classification based on ML depends on the mathematical structure of the classifier. For this thesis, we have selected a DNN and logic-based classifier for the purposes of classification. For the DNN case we have utilised the state-of-the-art CNN, Recurrent Neural Network (RNN), a combination of the CNN and RNN models, and Restricted Boltzmann Machine (RBM) techniques. We have employed a few logic-based algorithms, and made a detailed performance analysis of our selected data by the state-of-the-art XGBoost algorithm.

- **Performance Evaluation Methods:** The performance of a classifier model is always evaluated by some performance-measuring parameters. The performance of the utilised classifier models has been evaluated utilising evaluation criteria, namely Confusion Matrix, Sensitivity (Recall), Specificity, Precision, F-measure, False Positive Rate, False Negative Rate, Matthews Correlation Coefficient (M.C.C) and Mean-Square Error (MSE) value. Computational complexity and latency are big issues concerning the current DNN model. The performance of a few models has been evaluated based on the amount of time and parameters required to evaluate the model.
- **Feature-Selection Methods:** The identification of the relative importance of the features in the model can largely reduce the model complexity and computational time. The Feature-Selection method has been utilised to find the features which are more important to construct the classifier model. To reduce the data dimensionality and improve the computational latency, two Feature-Selection methods, Filter and Wrapper, have been utilised.
- **Utilised Platform:** We have performed all the simulation on a desktop computer with a Graphical Processing Unit (GPU) at the back end. All the simulations have been performed based on MATLAB, Python, TensorFlow, Keras and Weka environments.

1.3 Thesis Outline

This thesis is organised in a non-traditional "Thesis by publication" format. The format of this thesis has been approved by the Macquarie University Higher Degree Research (HDR) office. It consists of a general introduction and a conclusion, and the rest of most chapters is based on major scientific publications. Except for the introduction and conclusion chapters, all the texts and graphics of this thesis are either published or under review. The section headings have been retained as they are in the original publication. Figures, equations, tables, and references have been re-numbered and are in line with the thesis format.

The main objective of this thesis is to classify a set of Histopathological BC images into Benign and Malignant classes utilising a set of state-of-the-art ML techniques. The whole thesis is organised into ten chapters. Chapter 1 provides introductory information and an overall outline of the thesis, Chapter 2 provides a literature review, and Chapter 10 concludes the thesis with future directions. Chapter 3 to Chapter 9 are further classified based on DNN and logic-based algorithms as well as Global-Feature and Local-Feature extraction techniques. Figure 1.1 depicts how this thesis is organised chapter-wise.



1.4 Chapter-wise Contributions

Since this thesis follows the "Thesis by publication" format, except for the Introduction and Conclusion parts all the work in the other chapters has been published or is under review in a journal or conference publication. In the following we have summarised chapter-wise contributions and the corresponding publication information.

- Chapter 2 describes the involvement of ML techniques for BC image classification. Despite the importance of this challenging issue, to the best of the author's knowledge only a few literature review papers have been published; they have some strong findings as well as a few shortcomings. It is evident that most of the literature does not provide a holistic approach to the detailed procedures of BC image classification. This chapter provides a comprehensive approach to BC image-classification issues. This chapter also summarises the following:

1. For the ML techniques, especially in the classification task, a dataset or a set of datasets has been considered for the investigation. A few benchmark datasets are available, however, they are in a scattered state. In this chapter a summary table of most of the available breast-image datasets such as the Mammographic Image Analysis Society (MIAS) database, the Digital Database for Screening Mammography (DDSM), etc., are provided, with specifications.
2. Statistics have been summarised about recent findings based on the particular datasets.
3. Considering the importance of features for the classification task, this chapter summarises a few feature sets. Specifically this chapter categorises the features into local and global subsections. Later, local features are subdivided.
4. A detailed finding about BC image classification has been summarised with the sub-categories: a) Supervised b) Un-Supervised c) Semi-Supervised learning,

with the key contributions and shortcomings. Chapter 2 is covered by the following publication:

- I. A. A. Nahid and Y. Kong, “Involvement of Machine Learning for Breast Cancer Image Classification: A Survey”, *Computational and Mathematical Methods in Medicine, Hindawi*, pp. 1–29, vol. 2017.
- Images normally possess statistical and structural information. To take advantage of this statistical and structural information, clustering methods are utilised to cluster the image in an unsupervised way, and this image is then guided to the DNN model for classification purposes. Along with this, Chapter 3 proposes three novel DNN models for the image classification:
 1. First, one based on the CNN model.
 2. Second, one based on the LSTM model.
 3. Third, a combination of the CNN and LSTM models.

Except for the decision layers, almost all the layers in a DNN model are utilised, either for the extraction of the features globally or for removing features so that the complexity is reduced. However, at the end of the network a classifier layer has been used for the data classification. Usually a Softmax layer is used in the literature, however, in this model we have utilised both a Softmax layer and a Support Vector Machine (SVM) layer for the image classification, and compared their performance. Chapter 3 is covered by the following publication:

- II. A. A. Nahid, M. A. Mehrabi and Y. Kong, “Histopathological Breast-Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering”, *BioMed Research International, Hindawi*, pp. 1-20, 2018.

- In the CNN, a nonlinear function has been used to avoid repetition of the same behaviour. The most popular non-linear function is the Rectified Linear Unit (ReLU), which provides almost identical behaviour for all negative values. However, large and small negative values might not contain the same meaningful information. To avoid this issue and utilise the negative-value information we have implemented the Max-Min convolutional model for the image classification. Chapter 4 is covered by the following publication:

III. A. A. Nahid, F. B. Ali, and Y. Kong, “Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network”, *in 2017 20th International Conference on Computer and Information Technology (IC-CIT)*, *IEEE*, pp. 1-6, Dec. 2017.

- In Chapter 5, we have extracted handcrafted information such as Histogram and Local Binary Pattern (LBP) from the images. The CNN model is driven by both the raw images and the local features, and by utilising this information the CNN models have extracted the hierarchical global features to classify the images. Chapter 5 is covered by the following publication:

IV. A. A. Nahid and Y. Kong, “Local and Global Feature Utilization for Breast Image Classification by Convolutional Neural Network”, *in 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, *IEEE*, pp. 1-6, Nov. 2017

- In the ML techniques, a hypothetical model is created based on the learning from the available training data. This learning can be done from scratch or from a reference point. Learning from the training data is very important for the model to provide decisions. However, learning from scratch is not always as fruitful as

learning from references. A Recurrent Neural Network (RNN) has the property of learning from the reference point by feeding back information. The conventional RNN method suffers from a vanishing/exploding gradient problem. To avoid this issue we have utilised two methods, Long Short Term Memory (LSTM) and Gated Recurrent Units (GRUs). Frequency-domain data carry significant information. In Chapter 6, instead of utilising raw images directly, we have extracted frequency-domain information and so designed a novel classifier based on LSTM and GRU. Chapter 6 is covered by the following publication:

V. A. A. Nahid, M. A. Mehrabi and Y. Kong, “Frequency-Domain Information Along with LSTM and GRU Methods for Histopathological Breast-Image Classification”, in *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, pp. 1-6, Dec. 2017.

- The ResNet model is an advanced addition to CNN techniques. In Chapter 7 we have proposed two novel CNN models for Histopathological BC image classification. As an input we have utilised local features as well as frequency-domain features. As the edges of the images contain a significant amount of information, to extract the smooth-edge information we have utilised the CT as input for the image classifier. Chapter 7 is covered by the following publication:

VI. A. A. Nahid and Y. Kong, “Histopathological Breast-Image Classification Using Local and Frequency Domains by Convolutional Neural Network”, *Information, MDPI*, vol. 9, no. 1, pp. 1–26, 2018.

- Chapter 8 utilises an unsupervised Restricted Boltzmann Machine (RBM) for the BC Histopathological image classification, guided by supervised back propagation techniques. For the back propagation, Scaled Conjugate Gradient techniques have been adopted and as a feature we have utilised Tamura features.

Chapter 8 is covered by the following publication:

VII. A. A. Nahid, A. Mikaelian and Y. Kong, “Histopathological Breast Image Classification with Restricted Boltzmann Machine along with Back Propagation”, *BioMed Research, Allied Academics*. [Accepted on 2nd April, 2018].

- In Chapter 9 we have utilised local features such as Histogram, LBP, Harlick and Tamura features and utilised a few logic-based algorithms, especially XGBoost, for Breast-image classification. A classifier’s computational complexity and timing latency depends on the required number of parameters and features utilised. To reduce the number of parameters, too many of which increase the training time, and to enhance generality, a few Feature-Selection methods have been developed. This feature-reduction method can also be useful for reduction of the data dimensionality.

Chapter 9 is covered by the following publication:

VIII. A. A. Nahid, A. Mikaelian, M. A. Mehrabi and Y. Kong, “Histopathological Breast-Cancer Image Classification with Feature Prioritisation”, *BioMed Research International, Hindawi*. [in review].

1.5 Author’s Contributions

The major investigation, model designing, data processing, writing, drafting and editing has been done by myself (Abdullah-Al Nahid (AAN)) with invaluable guidance and suggestions provided by my principal supervisor Dr Yinan Kong (YK). In my PhD journey Mohamad Ali Mehrabia (MAM), Aaron Mikaelian (AM) and Ferdous Bin Ali (FA) also provided support with the writing up and presentation of papers. Table 1.2 provides details of the contributions made by various authors to the papers.

Table 1.2: Individual Contributions

Scope	Paper ID							
	I	II	III	IV	V	VI	VII	VIII
Data Collection	AAN, YK	AAN, YK	AAN, YK	AAN, YK	AAN, YK	AAN, YK	AAN, YK	AAN, YK
Paper Sturcture and Design	AAN	AAN	AAN	AAN	AAN	AAN	AAN	AAN
Concept Model Design	AAN	AAN	AAN	AAN	AAN	AAN	AAN	AAN
Analysis and Interpretation	AAN	AAN	AAN	AAN	AAN	AAN	AAN	AAN
Article Writing	AAN	AAN	AAN	AAN	AAN	AAN	AAN, AM	AAN, AM
Article Editing	AAN, YK	AAN, MAM, YK	AAN, FA, YK	AAN, YK	AAN, YK	AAN, YK	AAN, AM, YK	AAN, AM, YK
Overall Responsibility	YK	YK	YK	YK	YK	YK	YK	YK

Chapter 2

Involvement of Machine Learning for Breast-Cancer Image Classification: A Survey

2.1 Abstract

Breast-Cancer is one of the largest causes of womens' death in the world today. Advance engineering of natural image classification techniques and Artificial Intelligence methods has largely been used for the breast-image classification task. The involvement of digital image classification allows the doctor and the physicians a second opinion, and it saves the doctors' and physicians' time. Despite the various publications on breast image classification, very few review papers are available which provide a detailed description of Breast-Cancer image-classification techniques, Feature-Extraction and selection procedures, classification measuring parameterizations as well as image classification findings.

Published as: A. A. Nahid, Y. Kong, "Involvement of Machine Learning for Breast Cancer Image Classification: A Survey", *Computational and Mathematical Methods in Medicine, Hindawi*, pp. 1–29, vol. 2017.

We have put a special emphasis on the Convolutional Neural Network (CNN) method for breast image classification. Along with the CNN method we have also described the involvement of the conventional Neural Network (NN), Logic Based classifiers such as the Random Forest (RF) algorithm, Support Vector machines (SVM), Bayesian methods, and a few of the Semi-supervised and Unsupervised methods which have been used for breast image classification

2.2 Introduction

The Cell of the body maintain a cycle of regeneration processes. The balanced growth and death rate of the cells normally maintain the natural working mechanism of the body, but this is not always the case. Sometimes an abnormal situation occurs, where a few cells may start growing aberrantly. This abnormal growth of cells creates cancer, which can start from any part of the body and be distributed to any other part. Different types of cancer can be formed in human body; among them breast cancer creates a serious health concern. Due to the anatomy of the human body, women are more vulnerable to breast cancer than men. Among the different reasons for breast cancer, age, family history, breast density, obesity and alcohol intake are reasons for breast cancer.

Statistics reveal that in the recent past the situation has become worse. As a case study, Figure 2.1 shows the breast cancer situation in Australia for the last 12 years. This figure also shows the numbers of new males and females to start suffering from breast cancer. In 2007, the number of new cases for breast cancer was 12775, while the expected number of new cancer patients in 2018 will be 18235. Statistics show that in the last decade, the number of new cancer disease patients increased every year at an alarming rate.

Figure 2.2 shows the number of males and females facing death due to breast cancer.

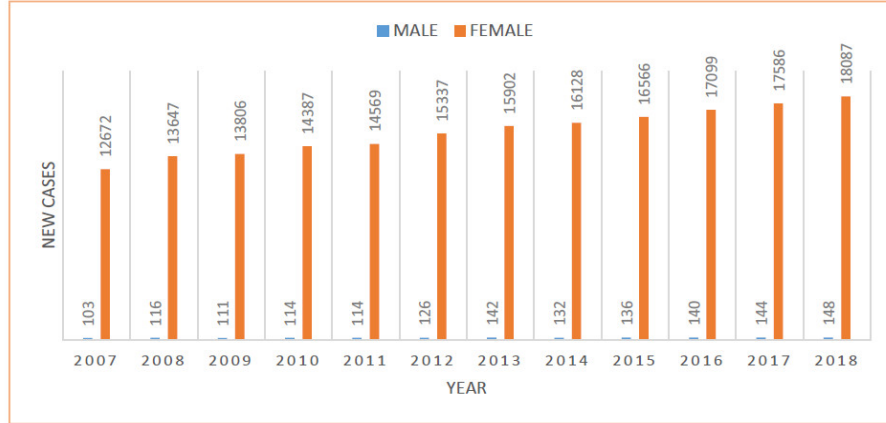


Figure 2.1: Numbers of new people facing cancer in Australia from 2007 to 2018 [1].

It is predicted that in 2018 around 3156 people will face death; among them 3128 will be women which is almost 99.11% of the overall deaths due to breast cancer.

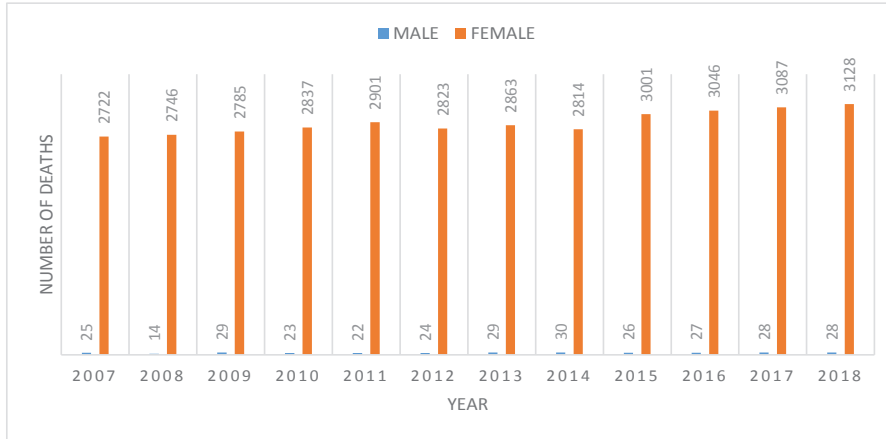


Figure 2.2: Numbers of people dying due to cancer in Australia from 2007 to 2018 [1].

Womens' breasts are constructed by lobules, ducts, nipples and fatty tissues. Milk is created in lobules and carried towards nipple by ducts. Normally epithelial tumors grows inside lobules as well as ducts and later form cancer inside the breast [3]. Once the cancer has started it also spreads to other parts of the body. Figure 2.3 shows the internal construction from a breast image.

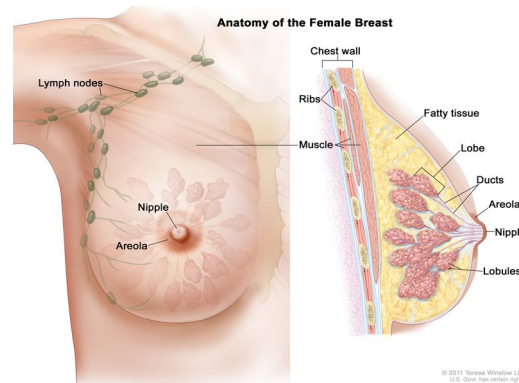


Figure 2.3: Anatomy of the female breast images (For the National Cancer Institute © 2011; Terese Winslow, U.S. Govt, has certain rights)

Breast Cancer tumors can be categorised into two broad scenarios:

- Benign (non-cancerous):

Benign cases are considered as non-cancerous, that is non-life threatening. But on a few occasions it could turn into a cancer status. An immune system known as "sac", normally segregates benign tumors from other cells, and can be easily removed from the body.

- Malignant (cancerous):

Malignant cancer starts from an abnormal cell growth, and might rapidly spread or invade nearby tissue. Normally the nuclei of the malignant tissue are much bigger than in normal tissue, which can be life threatening in future stages.

Cancer is always a life threatening disease. Proper treatment of cancer saves peoples lives. Identification of the normal, benign and malignant tissues are very important steps for further treatment of cancer. For the identification of benign and malignant conditions, imaging of the targeted area of the body helps the doctor and the physician for further diagnosis. With the advanced of modern photography techniques, the image of the targeted part of the body can be captured more reliably. Based on the penetration

of the skin and damage of the tissue medical photography techniques can be classified into two groups:

- Non-invasive:
 - a. Ultrasound: This photography technique uses similar techniques to SOund Nav-igation And Ranging (SONAR) which operates in the very-high-frequency domain and records the echos of that frequency, invented by Karl Theodore Dussik [4]. An ultrasound image machine contains a Central Processing Unit (CPU), transducer, a display unit and a few other peripheral devices. This device is capable of capturing both 2-D and 3-D images. Ultrasound techniques do not have any side-effects, with some exceptions like production of heat bubbles around the targeted tissue.
 - b. X-Ray : X-rays utilize electromagnetic radiation, invented by Wilhelm Conrad Roentgen in 1895. The Mammogram is a special kind of X-ray (low-dose) imaging technique which is used to capture a detailed image of the breast [5]. X-rays sometimes increase the hydrogen peroxide level of the blood, which may cause cell damage. Sometimes X-Rays may change the base of DNA.
 - c. Computer Aided Tomography (CAT) : CAT, or in short CT imaging, is advanced engineering of X-ray imaging techniques, where the X-ray images are taken at different angles. The CT imaging technique was invented in 1970 and has been mostly used for three dimensional imaging.
 - d. Magnetic Resonance Imaging (MRI): MRI is a non-invasive imaging techniques which produces a 3D image of the body, invented by Professor Sir Peter Marsfield, and this method utilizes both a magnetic field as well as radio waves to capture the images [6]. MRI techniques take longer to capture images, which may create discomfort for the user. Extra cautions need to be addressed to patients who may have implanted extra metal.
- Invasive:

a. Histopathological Images (Biopsy Imaging): Histopathology is the microscopic investigation of a tissue. For histopathological investigation, a patient needs to go through a number of surgical steps. The photographs taken from the histopathological tissue provide histopathological images.

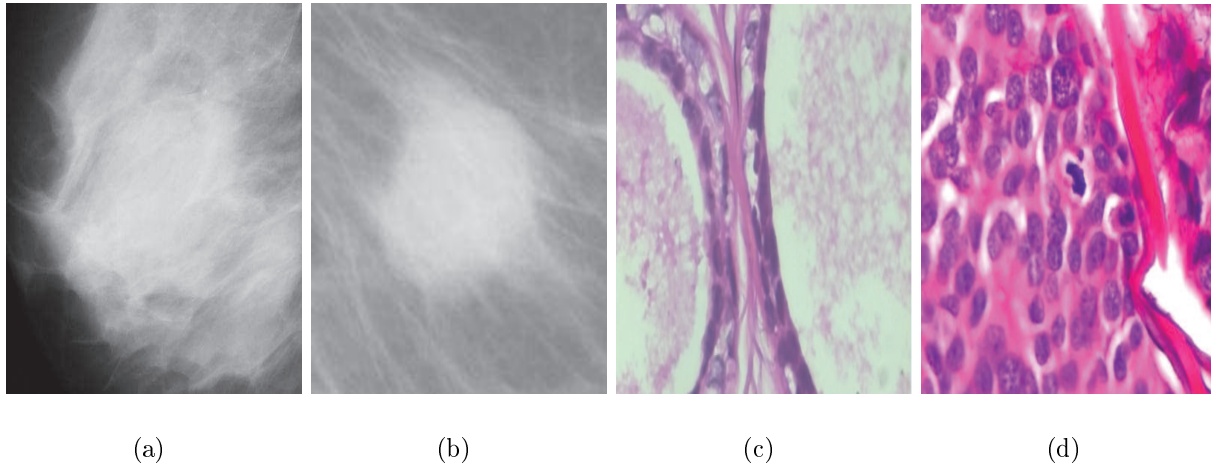


Figure 2.4: (a,b) show mammograms of benign and malignant images (Examples of non-invasive image) and (c,d) show histopathological benign and malignant images (Examples of invasive image)

2.3 Breast-Image Classification

Various algorithms and investigation methods have been used by researchers to investigate breast images from different perspectives depending on the demand of the disease, the status of the disease and the quality of the images. Among the different tasks, for breast image classification, Machine Learning (ML) and the Artificial Intelligence (AI) are heavily utilized. A general breast image classifier consists of the following four stages:

- Selection of a Breast Database
- Feature extraction and selection

- Classifier Model
- Performance Measuring Parameter
- Classifier Output.

Figure 2.5 shows a very basic Breast Image Classifier model.



Figure 2.5: A very basic Breast-Image Classification Model

2.3.1 Available Breast-Image Databases

Doctors and physicians are heavily reliant on the Ultrasound, MRI, X-Ray etc. images to find the breast cancer present status. However, to ease the doctors' work, some research groups are investigating how to use computers more reliably for breast cancer diagnostics. To make a reliable decision about the cancer outcome, researchers always base their investigation on some well- established image database. Various organizations have introduced sets of images databases which are available to researchers for further investigation. Table 2.1 gives a few of the available image databases, with some specifications.

Table 2.1: Available Breast-Image Database for Biomedical Investigation

Database	Number of		Database Size (GB)	Image Capture Technique		Image Type	Total Patients
	Images						
MIAS	322		2.3	Mammogram			161
DDSM				Mammogram			2620
CBIS-DDSm	4067		70.5	MG	DICOM		237
ISPY1	386,528		76.2	MR,SEG			237
Breast-MRI-NACT-Pilot	99,058		19.5	MRI			64
QIN-Breast	100835		11.286	PET/CT,MR	DICOM		67
Mouse-Mammary	23487		8.6	MRI	DICOM		32
TCGA-BRCA	230167		88.1	MR, MG	DICOM		139
QIN Breast DCE-MRI	76328		15.8	CT	DICOM		10
BREAST-DIAGNOSIS	105050		60.8	MRI/PET/CT	DICOM		88
RIDER Breast MRI	1500		.401	MR	DICOM		5
BCDR				Mammogram			1734
TCGA-BRCA			53.92(TB)	Histopathology			1098
BreakHis	7909			Histopathology			82
Inbreast	419			Mammogram			115

The image formats of the different databases are different. Few of the images contained images in JPEG format, few databases contained DICOM-format data. Herethe MIAS, DDSM and Inbreast databases contains mammogram images. According to the Springer (www.springer.com), Elsevier (www.elsevier.com), IEEE (www.ieeexplore.ieee.org) web sites, researchers have mostly utilized the MIAS and DDSM databases for the breast image classification research. The number of conference papers published for the DDSM and MIAS databases 110 and 168 respectively, with 82 journal papers published on DDSM databases and 136 journal papers have published using the MIAS database. We have verified these statistics on both Scopus (www.scopus.com) and the Web of Science database (www.webofknowledge.com). Figure 2.6 shows the number of published breast image classification papers based on the MIAS and DDSM data base from the year 2000 to 2017. Histopathological images provide valuable information and are being intensively

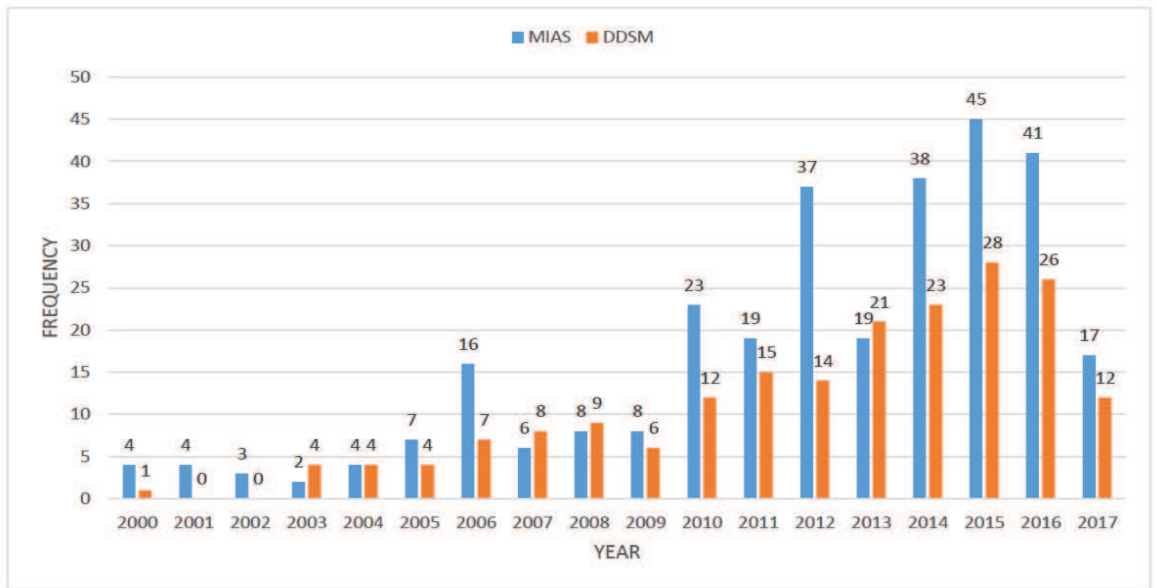


Figure 2.6: Number of papers published based on MIAS and DDSM databases

investigated by doctors for finding the current situation of the patient. The TCGA-BRCA and BreakHis databases contain histopathological images. Research has been performed

in a few experiments on this database too. Among these two database, BreakHis is the most recent histopathological image database, containing a total of 7909 images which have been collected from 82 patients [7]. So far around twenty research papers have been published based on this database.

2.3.2 Feature-Extraction and Selection

An important step of the image classification is extracting the features from the images. In the conventional image classification task, features are crafted in locally using some specific rules and criteria. However, the-state-of-the-art Convolutional Neural Network (CNN) techniques generally extract the features globally using kernels and these global features have been used for image classification.

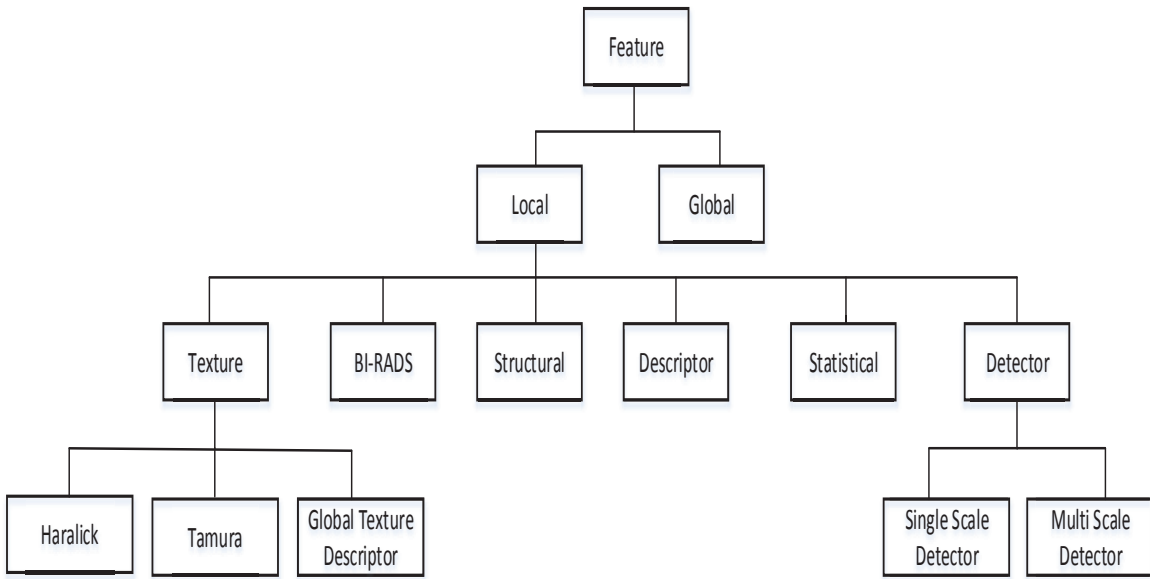


Figure 2.7: Classification of Features for Breast-Image Classification

Among the local features, Texture, Detector, Statistical are being accepted as important features for breast image classification. Texture features actually represent the low- level feature information of an image, which provides more detailed information of

an image that might be possible from histogram information alone. More specifically, texture features provide the structural and dimensional information of the color as well as the intensity of the image. Breast Imaging-Reporting and Data System (BI-RADS) is a mammography image assessment technique, containing 6 categories normally assigned by the radiologist. Feature detector actually provides information whether the particular feature is available in the image or not. Structural features provide information about the features structure and orientation such as the area, Convex Hull, Centroid. This kind of information give more detailed information about the features. In a cancer image, it can provide the area of the nucleus or the centroid of the mass. Mean, Median, Standard deviation always provide some important information on the dataset and their distribution. This kind of features has been categorized as statistical features. The total hierarchy of the image Feature-Extraction is resented in Figure 2.7. Table 2.2 and Table 2.3 further summarize the local features in detail.

Table 2.2: Feature Descriptor (Part 1)

Feature Category	Feature Description
Texture	Haralick texture features [8]
	(1) Angular Second Moment (ASM), (2) Contrast, (3) Correlation, (4) Sum of Squares of Variances (SSoV), (5) Inverse of Difference (IoD), (6) Sum of Average (SoA), (7) Sum of Variances (SoV), (8) Sum of Entropy (SoE), (9) Entropy, (10) Difference of Variance (DoV), (11) Difference of Entropy (DoE), (12) Gray-Level Concurrence Matrix (GLCM).
	Tamura features [9]
	(1) Coarseness, (2) Contrast, (3) Directionality, (4) Line-likeness, (5) Roughness, (6) Regularity.
Detector	Global Texture Descriptor
	(1) Fractal Dimension (FD), (2) Coarseness, (3) Entropy, (4) Spatial Gray-Level Statistics (SGLS), (5) Circular Moran Autocorrelation Function (CMAF).
	Single-Scale Detector
Detector	(1) Moravec's Detector (MD) [10], (2) Harris Detector (HD) [11], (3) Smallest Univariate Segment Assimilating Nucleus (SUSAN), [12]
	(4) Features from Accelerated Segment Test (FAST) [13], [14], (5) Hessian Blob Detector (HBD) [15], [16].
	Multi-Scale Detector [9]
	(1) Laplacian of Gaussian (LoG) [10] [17], (2) Difference of Gaussian (DoG) Contrast [18] (3) Harris Laplace (HL), (4) Hessian Laplace (HeL), (5) Gabor-Wavelet Detector (GWD) [19].
Strutural	(1) Area, (2) Bounding Box, (3) Centroid, (4) Convex Hull (CH), (5) Eccentricity, (6) Convex Image (CI), (7) Compactness, (8) Aspect Ratio (AR), (9) Moments, (10) Extent, (11) Extrema, (12) Major Axis Length (MaAL), (13) Minor Axis Length (MiAL),
	(14) Maximum Intensity (MaI), (15) Minimum Intensity (MiI), (16) Mean Intensity (MI), (17) Orientation, (18) Solidity.

Table 2.3: Feature Descriptor (Part 2)

Feature Category	Feature Description
Statistical	(1) Mean, (2) Median, (3) Standard Deviation, (4) Skewness, (5) Kurtosis, (6) Range, (7) Median.
	(1) Scale Invariant Feature Transform (SIFT) [18] [20], (2) Gradient Location-Orientation Histogram (GLOH) [21], (3) Speeded-Up Robust Features Descriptor (SURF) [22], [23], [24], (4) Local Binary Pattern (LBP), [25] [26] [27], [28], (5) Binary Robust Independent Elementary Features (BRIEF) [29], (6) Weber Local Descriptor (WLD) [30] [31] (7) Back Ground Local Binary Pattern (BGLBP) [32], (8) Center-Symmetric Local Binary Pattern (CS-LBP), [33] (9) Second-order Center-Symmetric Local Derivative Pattern (CS-LBP) [34], (10) Center-Symmetric Scale Invariant Local Ternary Patterns (CS-SILTP) [35], (11) Extended LBP or Circular LBP(E-LBP) [36], (12) Opponent Color Local Binary Pattern (OC-LBP) [37], (13) Original LBP(O-LBP) [26], (14) Spatial extended Center-Symmetric Local Binary Pattern (SCS-LBP) [38], (15) Scale Invariant Local Ternary Pattern (SI-LTP) [39], (16) Variance-based LBP (VAR-LBP) [25], (17) eXtended Center-Symmetric Local Binary Pattern (XCS-LBP), (18) Average Local Binary Pattern (ALBP), (19) Block Based Local Binary Pattern (BBLBP) [40],
Descriptor	(1) Margin Integrality (MarI), (2) Margin Ambiguity (MarA), (3) Echo Pattern Posterior Feature(EPPF), (4) Calcification in Mass (CM), (5) Architectural Distortion (AD), (6) Edema, (7) Eymph Nodes Axillary (ENA) (8) Ducts changes(DC), (9) Skin thickening (ST), (10) Post Surgical Fluid Collection (PSFC), (11) Skin Retraction(SR1), (12) Fat Necrosis (FN), (13) Lump nodes Intramammary (LNI).

BI-RADS [41]

Features which are extracted for classification do not always carry the same importance. Some features may even contribute to degrading the classifier performance. Prioritization of the feature set can reduce the classifier model complexity and so it can reduce the computational time. Feature set selection and prioritization can be classified into three broad categories:

- Filter: The filter method select features without evaluating any classifier algorithm.
- Wrapper: The wrapper method selects the feature set based on the evaluation performance of a particular classifier.
- Embedded: The embedded method takes advantage of the filter and wrapper methods for classifier construction.

Figure 2.8 shows a generalized feature selection method where we have further classified the filter method into Fisher Score, Mutual Information, Relief and Chi Square methods. The embedded method has been classified in to Bridge Regularization, Lasso and Adaptive Lasso methods, while the wrapper method has been classified to Recursive Feature Selection and Sequential Feature Selection method.

2.3.3 Classifier Model

Based on the learning point of view, breast image classification techniques can be categorized into the following three classes [42]:

- Supervised
- Un-supervised
- Semi-Supervised

These three classes can be split into Deep Neural Network (DNN) and Conventional Classifier (Without DNN) and to some further classes as in Table 2.4.

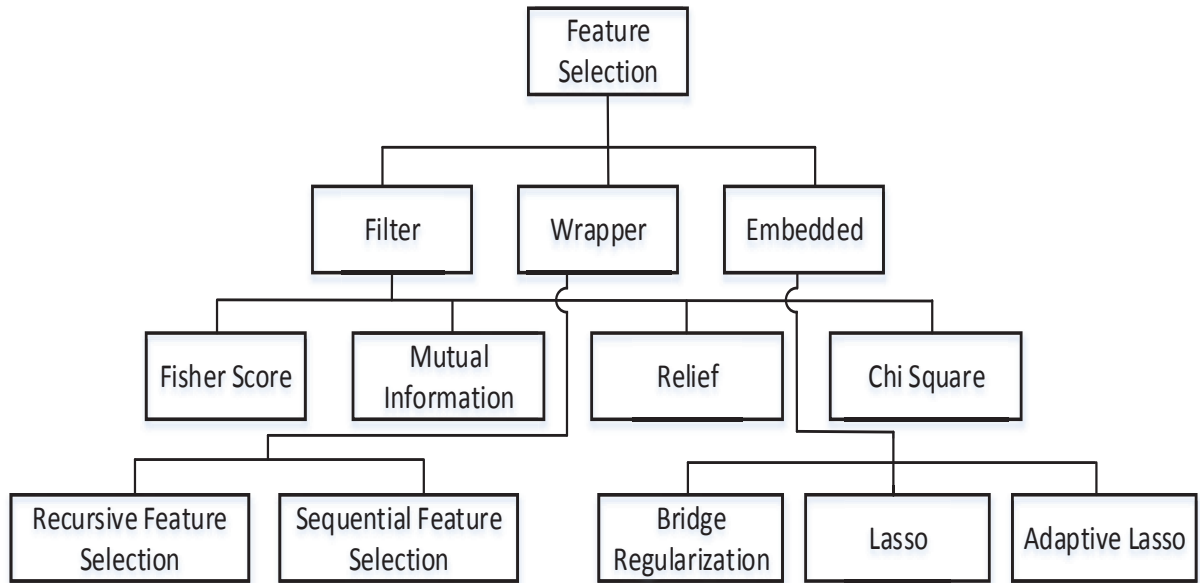


Figure 2.8: A summary of Feature-Selection Method

2.3.4 Performance-Measuring Parameters

A Confusion Matrix is a two-dimensional table which is used to give a visual perception of classification experiments [43]. The $(i, j)^{\text{th}}$ position of the confusion table indicates the number of times that the i^{th} object is classified as the j^{th} object. The diagonal of this matrix indicates the number of times the objects are correctly classified. Figure 7.9 shows a graphical representation of a Confusion Matrix for the binary classification case.

Hypothesized Class	
True Class	True Positive (A)
	False Negative (B)
True Class	False Positive (C)
	True Negative (D)

Figure 2.9: Confusion Matrix

Among the different classification performance properties, this matrix will provide following parameters:

- Recall is defined as $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$.
- Precision is defined as: $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$.
- Specificity is defined as: $\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$.
- Accuracy is defined as $\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$.
- F-1 score is defined as $F_1 = \frac{2 \times \text{Recall}}{2 \times \text{Recall} + \text{FP} + \text{FN}}$.
- Matthew Correlation Coefficient (MCC): MCC is a performance parameter of a binary classifier, in the range $\{-1 \text{ to } +1\}$. If the MCC values tend more towards +1, the classifier gives a more accurate classifier, the opposite condition will occur if the value of the MCC tend towards the -1 . MCC can be defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (2.1)$$

2.4 Performance of different Classifier models on Breast-Image Dataset

Based on Supervised, Semi-Supervised and Un-Supervised methods different research groups have been performed classification operation on different image data base. In this section we have summarized few of the works of breast image classification.

2.4.1 Performance Based on Supervised Learning

In supervised learning, a general hypothesis is established based on externally supplied instances to produce future prediction. For the supervised classification task, features are

extracted or automatically crafted from the available data set and each sample is mapped to a dedicated class. With the help of the features and their levels a hypothesis is created. Based on the hypothesis unknown data are classified [44].

Figure 2.10 represents an overall supervised classifier architecture. In general, the whole data-set is split into training and testing parts. To validate the data, some time data are also split into a validation part as well. After the data splitting the most important part is to find out the appropriate features to classify the data with the utmost accuracy. Finding the features can be classified into two categories, Locally and Globally crafted. Locally crafted means that this method requires a hand-held exercise to find out the features, whereas globally crafted means that a kernel method has been introduced for the Feature-Extraction. Hand-crafted features can be prioritized, whereas global feature selection does not have this luxury.

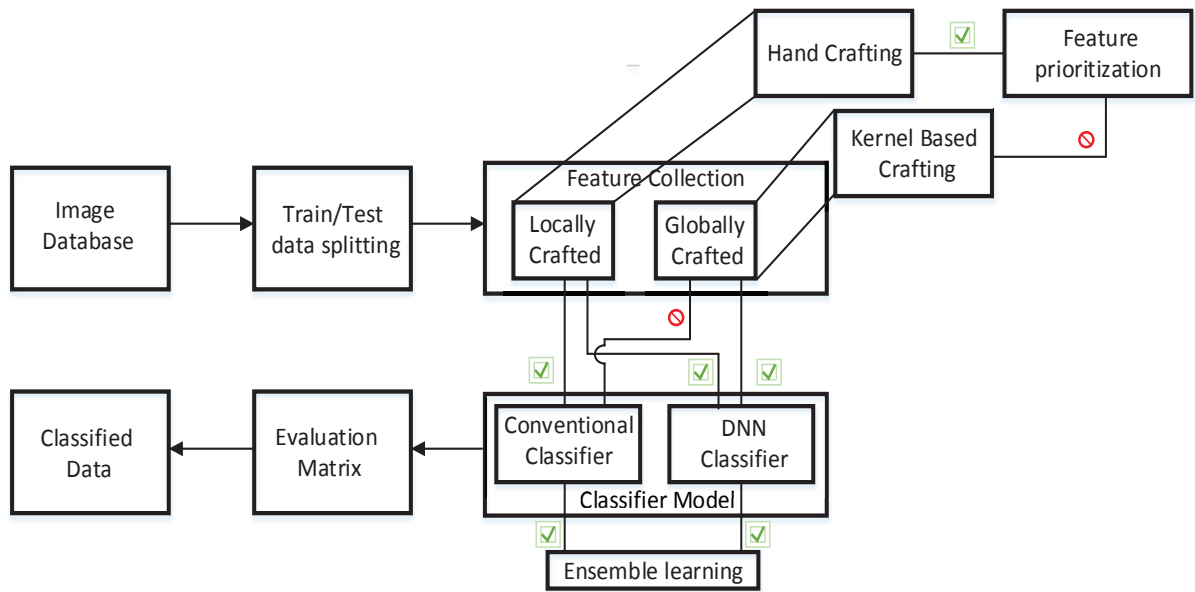


Figure 2.10: A generalised Supervised Classifier Model

Conventional Neural Network

The Neural Network (NN) concept comes from the working principle of the human brain. A biological neuron consists of the following four parts:

- Dendrites • Nuclease • Cell Body • Axon

Dendrites collect signals, Axons carry the signal to the next dendrite after processing by the cell body as the Figure 2.11. Using the neuron working principle, the perceptron

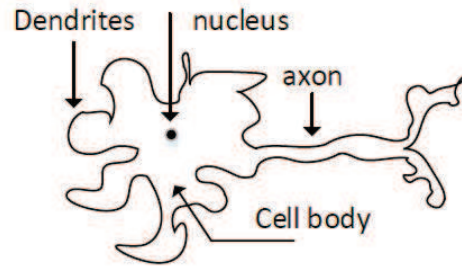


Figure 2.11: A model of a Biological neuron

model was proposed by Rosenblatt in 1957 [45]. A single-layer perceptron linearly combines the input signal and gives a decision based on a threshold function. Based on the working principle and with some advanced mechanism and engineering, NN methods have established a strong footprint in many problem-solving issues. Figure 2.12 shows the basic working principle of NN techniques.

In the NN model the input data $\mathbf{X} = \{x_0, x_1, \dots, x_N\}$ is first multiplied by the weight data $\mathbf{W} = \{w_0, w_1, \dots, w_N\}$ and then the output is calculated using

$$\mathbf{Y} = \mathbf{g}\left(\sum\right) \quad \text{where } \sum = \mathbf{W} \cdot \mathbf{X} \quad (2.2)$$

Function \mathbf{g} is known as the activation function. This function can be any threshold value or sigmoid or hyperbolic, etc. In the early stages, feed-forward neural network techniques were introduced [46], lately the backpropagation method has been invented to utilize the error information to improve the system performance [47], [48].

The history of breast image classification by NN is a long one. To the best of my knowledge much the pioneer work was performed by Dawson et al. in 1991 [49]. Since then, NN has been utilized as one of the strong tools for breast image classification. We have summarized some of the work related to NN and Breast image classification in Tables 2.5, 2.6 and 2.7.

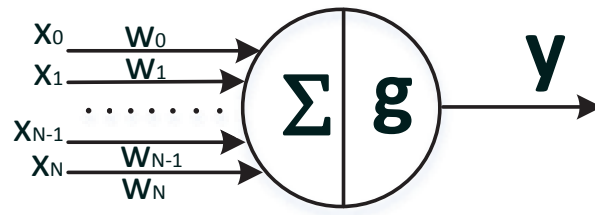


Figure 2.12: Working principle of a simple Neural Network technique

Table 2.4: A simplified Hierarchy of Classification

Learning Technique		Algorithm
Supervised	Conventional	(1) ID3, (2) C4.5, 3 (3) Bagging, (a) Logic Based $\left\{ \begin{array}{l} (4) \text{ Random Trees, } (5) \text{ Random Forest,} \\ (6) \text{ Boosting, } (7) \text{ Advanced Boosting,} \\ (8) \text{ Extreme Boosting (XGBoosting).} \end{array} \right.$
		(b) Bayesian $\left\{ \begin{array}{l} (1) \text{ Naive Bayes} \\ (2) \text{ Bayesian Network} \end{array} \right.$
		(c) Conventional Neural Network
		(d) Support Vector Machine
	DNN Based	(a) Convolutional Neural Network (CNN), (b) Deep Belief Network (DBN), (c) Generative Adversarial Network (GAN).
Un-Supervised	Conventional	(a) k-Means Clustering (b) Self Organizing Map (SOP) c) Fuzzy C-Means Clustering (FCM)
	DNN Based	(a) Deep Belief Network (DBN)
Semi-Supervised	Conventional	(a) Self Training (b) Graph Based (c) S3V3 (d) Multi-View (e) Generative model

Table 2.5: Neural Network for Breast-Image Classification (Part 1)

Reference	Descriptor	Image Type	No. of Images	Key Finding
Rajakeerthana et al. [50]	(1) GLCM, GLDM, SRDM, NGLCM, GLRM	Mammogram	322	(1) The classifier achieved 99.20% accuracy.
Lessa et al. [51]	(1) Mean, Median, Standard Deviation, Skewness, Kurtosis, Entropy, Range Median	Thermographic	94	(1) Achieved Sensitivity, Specificity and Accuracy are 87.00%, 83.00% and 85.00% respectively.
Wan et al. [52]	(1) ALBP (2) BBLBP	OCM	46	(1) Achieved Sensitivity and Specificity are 100% and 85.20% respectively. (2) ROC value obtained 0.959.
Chen et al. [41]	(1) 19 BI-RADS features have been used.	Ultra Sound	238	(1) Chi squared method has been utilized for the feature selection. (2) Achieved Accuracy, Sensitivity and Specificity are 96.10%, 96.70% and 95.70%. respectively.
Lima et al. [53]	(1) Total 416 features have been used.	Mammogram	355	(1) Multiresolution wavelet and Zernike moment have been utilized for the Feature-Extraction.
Albirami et al. [54]	(1) 12 Statistical measures such as Mean, Median, Max etc have been utilized as the features.	Mammogram	322	(1) Wavelet transform has been utilized for the Feature-Extraction. (2) The achieved Accuracy, Sensitivity and Specificity are 95.50%, 95.00% and 96.00% respectively.
El Atlas et al. [55]	(1) 13 Morphological features have been utilized.	Mammogram	410	(1) Firstly the edge information has been utilized for the mass segmentation and then the morphological features extracted. (2) Achieved best accuracy is 97.5 %

Table 2.6: Neural Network for Breast-Image Classification (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Finding
Alharbi et al. [56]	(1) 49 features have been utilized.	Mammogram	1100	(1) Five feature selection methods: Fisher score, Minimum Redundancy-Maximum Relevance, Relief-f, Sequential Forward Feature Selection, and Genetic Algorithm has been used.
				(2) Achieved Accuracy, Sensitivity and specificity are 94.20 %, 98.36 % and 99.27 % respectively
Peng et al. [57]	(1) Haralick and Tamura features have been utilized.	Mammogram	322	(1) Feature reduction has been performed by Rough-Set theory and selected 5 prioritized features.
				(2) The best Accuracy, Sensitivity and Specificity achieved was 96.00 %, 98.60 % and 89.30 %
Jalalian et al. [58]	(1) GLCM (2) Compactness.	Mammogram		(1) The obtained classifier Accuracy, Sensitivity, Specificity are 95.20 %, 92.40 % and 98.00 % respectively.
				(1) 2D contour of breast mass in mammography has been converted into 1D signature. (2) NN techniques achieved Accuracy is 99.60% when RMS slope utilized.
Chen et al. [60]	(1) Autocorrelation features	Ultrasound	242	(1) The overall achieved Accuracy, Sensitivity and Specificity is 95.00% 98.00% and 93% respectively.
Chen et al. [61]	(1) Autocorrelation features.	Ultrasound	1020	1. The obtained ROC area is 0.9840 ± 0.0072 .

Table 2.7: Neural Network for Breast-Image Classification (Part 3)

Reference	Descriptor	Image Type	No. of Images	Key Finding
Chen et al. [62]	(1) Variance contrast of Wavelet Coefficient	Ultrasound	242	(1) The achieved ROC curve 0.9396 ± 0.0183
	(2) Auto correlation of Wavelet Coefficient.			
	(1) 22 different morphological features such as convexity, lobulation etc. have been utilized.			
	(1) Age of Patient, (2) Mass Shape, (3) Mass border, (4) Mass density, (5) BIRADS.			
Silva et al. [63]	(1) 22 different morphological features such as convexity, lobulation etc. have been utilized.	Ultrasound	—	(1) The best obtained Accuracy and ROC curve are 96.98 % and 0.98 respectively
Saritas [64]	(1) Age of Patient, (2) Mass Shape, (3) Mass border, (4) Mass density, (5) BIRADS.	Mammogram	—	(1) Disease prediction rate is 90.5% (2) Neural Network utilized 5 neurons in input layers and one hidden layer.
Melendez et al. [65]	(1) Area, Perimeter etc. have been utilized. .	Mammogram	322	(1) The achieved Sensitivity and Specificity are 96.29% and 99.00% respectively.

Deep Neural Network

Deep Neural Network (DNN) is a state-of-the-art concept where conventional NN techniques have been utilized with advanced engineering. It is found that conventional NNs have difficulties in solving complex problems, where as DNN solve them with utmost precision. However DNNs suffer from more time and computational complexity than the conventional NN.

- Convolutional Neural Network (CNN)
- Deep Belief Network (DBN)
- Generative Adversarial Network (GAN)
- Recurrent Neural Network (RNN)

Convolutional Neural Network

A CNN model is the combination of a few intermediate mathematical structures. This intermediate mathematical structure creates or helps to create different layers:

- **Convolutional Layer**

Among all the other layers, the convolutional layer is considered as the most important part for a CNN model and can be considered as the backbone of the model. A kernel of size $m \times n$ is scanned through the input data for the convolutional operation which ensures the local connectivity and weight sharing property.

- **Stride and Padding**

In the convolutional operation, a filter scans through the input matrices. In each step how much position a kernel filter moves through the matrix is known as the stride. By default stride keeps to 1. With inappropriate selection of the stride the model can lose the border information. To overcome this issue the model utilize extra rows and columns at the end of the matrices, and this added rows and columns contain all 0s. This adding of extra rows and columns which contain only zero value is known as zero padding.

- **Nonlinear Operation**

The output of each of the kernel operations is passed through a rectifier function such as Rectified Linear Unit (ReLU), Leaky ReLU, TanH, Sigmoid etc. The Sigmoid function can be defined as

$$\sigma(x) = \frac{1}{(1 + \exp^{-x})} \quad (2.3)$$

and the Tanh function can be defined as

$$\tanh(x) = \frac{(\exp^x - \exp^{-x})}{(\exp^x + \exp^{-x})}. \quad (2.4)$$

However the most effective rectifier is ReLU. The ReLU method converts all the information into zero if it is less than or equal to zero, and passes all the other data as is as shown in Figure 5.4.

$$\sigma(x) = \max(0, x). \quad (2.5)$$

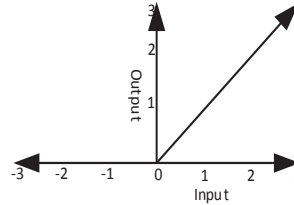


Figure 2.13: ReLU Operation

Another important nonlinear function is Leaky-ReLU

$$\text{Leaky-ReLU}(x) = \sigma(x) + \alpha \min(0, x) \quad (2.6)$$

where α is predetermined parameter which can be varied to give a better model.

- **Sub-Sampling**

Sub-sampling is the procedure of reducing the dimensionality of each of the feature

maps of a particular layer; this operation is also known as a pooling operation. Actually it reduces the amount of feature information from the the overall data. By doing so, it reduces the overall computational complexity of the model. To do this $s \times s$ patch units are utilized. The two most popular pooling methods are

- Max-Pooling
- Average Pooling.

In Max-Pooling, only the maximum values within a particular kernel size are selected for further calculation. Consider an example of a 16 x16 image as shown in Figure 5.4. A 2 by 2 kernel is applied to the whole image, 4 blocks in total, and produces a 4x4 output image. For each block of four values, we have selected the maximum. For instance from blocks one, two, three and four, maximum values 4, 40, 13, 8 are selected respectively as they are the maximum in that block. For the Average pooling operation, each kernel gives the output as average.

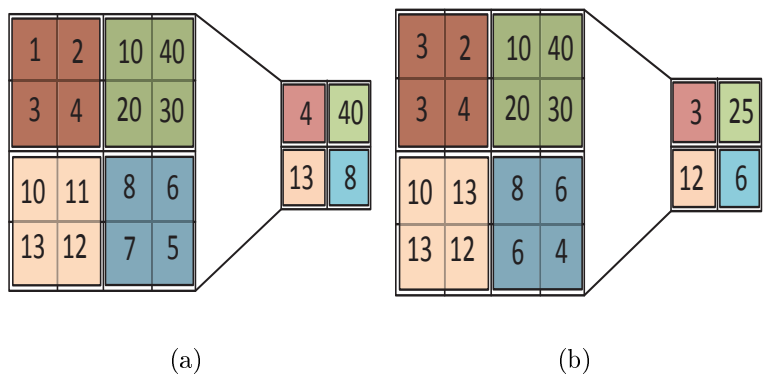


Figure 2.14: Max-Pooling and Average Pooling

• **Drop-Out**

Regularization of the weight can reduce the over-fitting problem. Randomly removing some neurons can regularize the over-fitting problem. The technique of randomly removing neurons from the network is known as dropout.

- **Soft-Max Layer**

This layer contains normalized exponential functions to calculate the loss function for the data classification.

Figure 7.7 shows a generalized CNN model for the image classification. All the neurons of the most immediate layer of a fully connected layer are completely connected with the fully connected layer, like a conventional Neural Network. Let f_j^{l-1} represent the j^{th} feature map at the layer $l - 1$. The j^{th} feature map at the layer l can be represented as

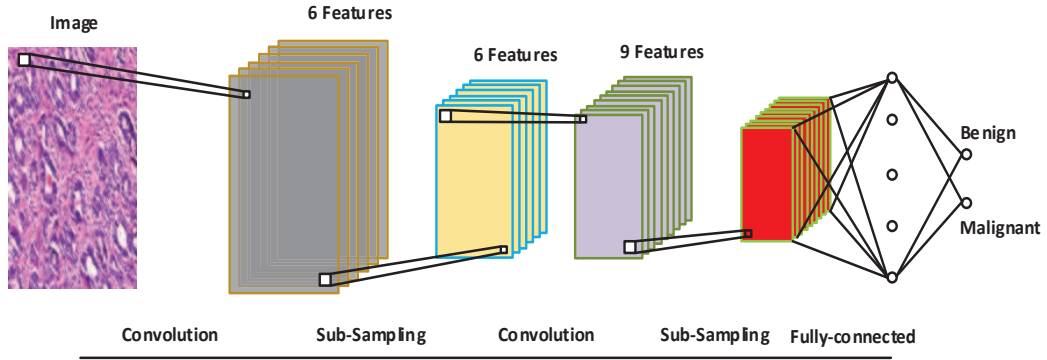


Figure 2.15: Workflow of a Convolutional Neural Network

$$f_j^l = \sigma\left(\sum_{i=1}^{N^{l-1}} f_i^{l-1} * k_{i,j} + b_j^l\right) \quad (2.7)$$

where N^{l-1} represents the number of feature maps at the $l - 1^{\text{th}}$ layer, $k_{i,j}$ represents the kernel function and b_j^l represents the bias at l , where σ performs a nonlinear function operation. The layer before the soft-max layer can be represented as

$$h_p^{\text{end}} = w^{\text{end}} * h_p^{\text{end}-1} + b^{\text{end}} \quad (2.8)$$

As we are working on a binary classification, the Soft-Max regression normalized output can be represented as

$$\bar{y}_p = \frac{\exp(h_p^{\text{end}})}{\sum_{p=1}^2 \exp(h_p^{\text{end}})} \quad (2.9)$$

Let $p = 1$ represent Benign class and $p = 2$ represents the Malignant class. The cross-entropy loss of the above function can be calculated as

$$L_p = -\ln(\bar{y}_p) \quad (2.10)$$

Whichever group experiences a large loss value, the model will consider the other group as predicted class.

A difficult part of working on DNN is that it requires a specialized software package for the data analysis. Few research groups have been working on how effectively data can be analyzed by DNN from different perspectives and the demand. Table 2.8 summarizes some of the software which is available for DNN analysis.

Table 2.8: Available Software for Deep-Learning Analysis

Software	Interface and Backend	Provider
Caffe [66], [67]	Python, MATLAB, C++	Berkeley Vision and Learning Centre, University of California, Berkeley
Torch [68]	C, LuaJIT	
MatConvNet [69], [70]	MATLAB, C	Visual Geometry Group, Department of Engineering, University of Oxford
Theano [71], [72]	Python	Montreal Institute for Learning Algorithms University of Montreal
TensorFlows [73]	C++, Python	Google
CNTK [74]	C++	Microsoft
Keras [75]	Theano, Tensor Flow	MIT
dl4j [76]	java	Skymin Engineering
DeeBNET [77], [78]	MATLAB	Information Technology Department, Amirkabir University of Technology

The history of the CNN and its use for biomedical image analysis is a long one. Fukushima first introduced a CNN named "neocognitron" which has the ability to recognize stimulus patterns with a few shifting variances [79]. To the best of our knowledge, Wu et al. first classified a set of mammogram images into malignant and benign classes using a CNN model [80]. In their proposed model they only utilized one hidden layer. After that, in 1996 Sahiner et al. utilized CNN model to classify mass and normal breast tissue and achieved ROC scores of 0.87 [81]. In 2002, Lo et al. utilized a Multiple Circular Path CNN (MCPCNN) for tumor identification from mammogram images and obtained ROC scores of around 0.89. After an absence of investigation of the CNN model, this model regained its momentum after the work of Krizhevsky et al [82]. Their proposed model is known as AlexNet. After this work a revolutionary change has been achieved in the image classification and analysis field. As an advanced engineering of the AlexNet, the paper titled "Going Deeper with Convolutions" by Szegedy [83] introduced the GoogleNet model. This model contains a much deeper network than AlexNet. Sequentially ResNet [84], Inception [85], Inception-v4, Inception-ResNet [86] and a few other models have recently been introduced.

Later, directly or with some advanced modification, these DNN models have been adapted for biomedical image analysis. In 2015, Fonseca et al. [87] classified breast density using CNN techniques. CNN requires a sufficient amount of data to train the system. It is always very difficult to find a sufficient amount of medical data for training a CNN model. A pre-trained CNN model with some fine tuning can be used rather than create a model from scratch [88]. The authors of [88] did not perform their experiments on a breast cancer image data set, however they have performed their experiments on three different medical data-sets with layer-wise training and claimed that "Retrained CNN along with adequate training can provide better or at least the same amount of performance".

The Deep Belief Network (DBN) is another branch of the Deep Neural Network, which

mainly consists of Restricted Boltzmann Machine (RBM) techniques. The DBN method was first utilized for supervised image classification by Liu [89]. After that, Zaher utilized the DBN method for breast image classification [90]. This field is still not fully explored for breast image classification yet. Zhang utilized both RBM and Point-Wise Gated RBM (PRBM) for shear-wave electrography image classification where the data set contains 227 images [91]. Their achieved classification Accuracy, Sensitivity, Specificity are 93.40%, 88.60% and 97.10% respectively. Table 2.9 and 2.11 summarize the most recent work for breast image classification along with some pioneer work on CNN.

Table 2.9: Convolutional Neural Network (Part 1)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Wu et al. [80]	(1) Global Features	Mammogram	40	(1) Achieved Sensitivity 75.00% and Specificity 75.00%.
Sabiner et al. [81]	(1) Global Features	Mammogram	168	(1) The achieved ROC score is 0.87.
Lo et al. [92]	(1) Density; Size, Shape, Margin.	Mammogram	144	(1) The achieved ROC curve is 0.89.
Fonseca et al. [87]	(1) Global Features	Mammogram	—	(1) Breast density classification has been performed utilizing HT-L3 convolution. (2) Average achieved obtained Kappa value is 0.58.
Arevalo et al. [93]	(1) Global Features.	Mammogram	736	(1) The achieved ROC curve is 0.826.
Su et al. [94]	(1) Global Features	Mammogram	92	(1) Fast Scanning CNN (fCNN) method has been utilized to reduce the information loss. (2) The average Precision, Recall and F1 score are 91.00%, 82.00% and 0.85 respectively.
Sharma et al. [95]	(1) GLCM, GLDM Geometrical	Mammogram	40	(1) The best Accuracy achieved is 75.23 % and 72.34 % respectively for fatty and dense tissue classification.
Spanhol et al. [7]	(1) Global Features	Histopathology	7909	(1) The best Accuracy achieved 89±6.6 %.
H. Rezaeilouyeh et al. [7]	1. Local and Global Features.	Histopathology	—	1. Shearlet Transform has been utilized for extracting local features. 2. When they utilize RGB image along with magnitude of Shearlet transform together, the Achieved Sensitivity, Specificity, Accuracy was 84.00±1.00%, 91.00±2.00% and 84.00±4.00%; when they utilize RGB image along with both the phase and magnitude of Shearlet transform together, the achieved Sensitivity, Specificity, Accuracy was 89.00±1.00%, 94.00±1.00% and 88.00±5.00%. "

Table 2.10: Convolutional Neural Network (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Albayrak et al. [96]	(1) Global Features	Histopathology	100	(1) Cluster-based segmentation has been performed to find out the cellular structure.
				(2) Blob analysis has been performed on the segmented images.
				(3) To reduce the high dimensionality, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods have been utilized.
				(4) Before the dimensionality reduction the Precision, Recall and F-score values were 97.20%, 66.00% and 0.78% respectively, but when the dimensionality reduction method was utilized the Precision, Recall and F-score values were 100.00%, 94.00% and 0.96% respectively
				(5) The best average accuracy is 73.00% (without dimensionality reduction) and 96.8 % (with dimensionality reduction).
				(1) They performed their experiments on the DDSM database.
Jiao et al. [97]	(1) Global and Local Features.	Mammogram	—	(2) Total required parameter is 5.8×10^7 and time for the per image processing is 1.10 ms.
				(3) The best classification achieved is 96.70%, however they show that when they utilize the VGG model the accuracy was 97.00% which is slightly better than their model. However in terms of memory size and time per image processing their model gives better performance than the VGG model.
Zejmo et al. [98]	(1) Global Features	Cytology	40	(1) GoogleNet and AlexNet models have been utilized.
				(2) The best accuracy obtained when they utilized GoogleNet model was 83.00%.

Table 2.11: Convolutional Neural Network (Part 3)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Jiang et al. [99]	(1) Global Features	Mammogram	—	(1) Image preprocessing was performed to enhance tissue characteristics. (2) Transfer learning was performed and obtained AUC 0.88 whereas when the system learned from scratch the best ROC is 0.82.
Suzuki et al. [100]	(1) Global Features	Mammogram	198	(1) The achieved sensitivity 89.90% (2) Transfer learning techniques has been utilized.
Qiu et al. [101]	1.Global Features	Mammogram	270	1. Average achieved accuracy is 71.40%.
Samala et al. [102]	(1) Global Features	—	92	(1) They utilized Deep Learning CNN (DLCNN) and CNN models for classification. 2. The AUC of CNN and DLCNN model is 0.89 and 0.93 respectively.
Sharma et al. [95]	1. Global Features	Mammogram	607	(1) Transfer learning and ensemble techniques utilized. (2) When using ensemble techniques the soft voting method has been used. (3) The best ROC score is 0.86.
Kooi et al. [103]	(1) Global and Local features	Mammogram	44090	(1) Transfer learning method utilized (VGG model).
Geras et al. [104]	(1) Global Features	Mammogram	102800	(1) They investigated the relation of the accuracy with the Database size and Image size.
Arevalo et al. [93]	(1) Global Features	Mammogram	736	(1) The best ROC value was 0.822.

Logic-Based Algorithm

A Logic-based algorithm is a very popular and effective classification method which follows the tree structure principle and logical argument. This algorithm classifies instances based on the feature's values. Along with other criteria, a decision-tree based algorithms contains the following features:

- Root Node: A root node contains no incoming node, and it may or may not contain any outgoing edge.
- Splitting: Splitting is the process of subdividing a set of cases into a particular group. Normally the following criteria are maintained for the splitting:
 - Information Gain
 - Gini Index
 - Chi-squared
- Decision Node
- Leaf/Terminal Node: This kind of node has exactly one incoming edge and no outgoing edge. The tree always terminate here with a decision.
- Pruning: Pruning is a process of removing subtrees from the tree. Pruning performs to reduce the over-fitting problem. Two kinds of pruning techniques are available:
 - Pre-Pruning
 - Post-Pruning

Among all the tree-based algorithms, Iterative Dichotomiser 3 (ID3) can be considered as a pioneer, proposed by Quinlan et al. [105]. The problem of the ID3 algorithm is to find the optimal solution which is very much prone towards over-fitting. To overcome the limitation of the ID3 algorithm the C4.5 algorithm has been introduced by Quinlan et al. [106], where a pruning method has been introduced to control the over-fitting problem. Pritom et al. [107] classified the Wisconsin breast dataset where they utilized 35 features. They have obtained 76.30% Accuracy, 75.10% False Positive Rate, ROC score 0.745 when

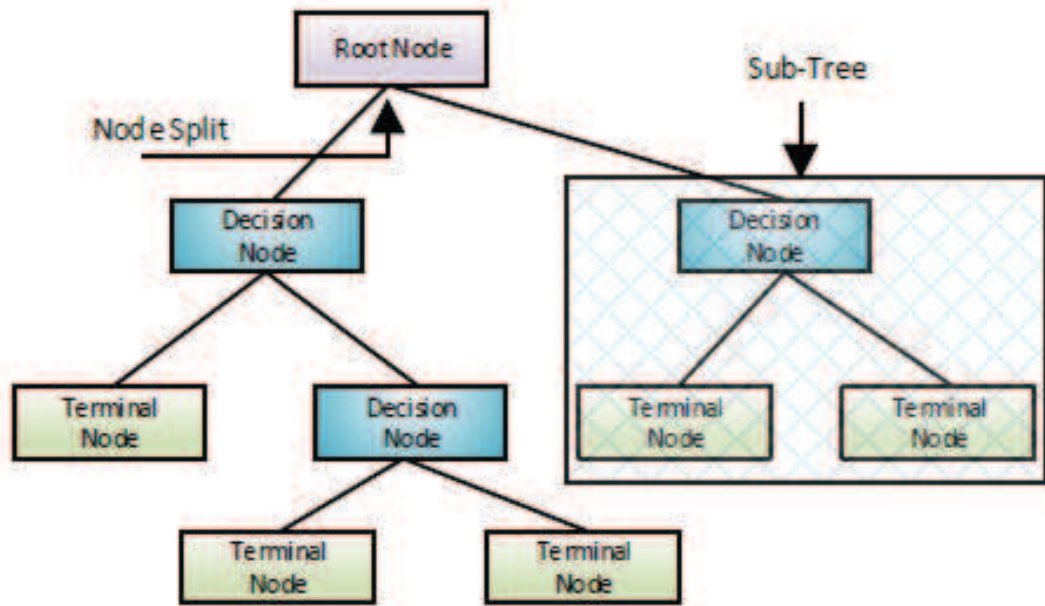


Figure 2.16: A general structure of a Tree

they ranked the features. Without ranking the features they obtained 73.70% Accuracy, 50.70% False Positive Rate, ROC score value 52.80. Asri et al. [108] utilized the C4.5 algorithm for the Wisconsin database classification where they utilized 11 features and obtained 91.13% Accuracy.

Logic-based algorithms allow us to produce more than one tree and combine the decisions of those trees for an advanced result; this mechanism is known as an ensemble method. An ensemble method combines more than one classifier hypothesis together and produces more reliable results through a voting concept. Boosting and bagging are two well-known ensemble methods. Both boosting and bagging aggregate the trees. The difference is, in bagging successive trees do not depend on the predecessor trees, where in the boosting method successive trees depend on the information gathered from the predecessor trees. Gradient boosting is a very popular method for data classification [109], [110], however a state-of-the-art Boosting algorithm such as " Extreme Gradient Boosting " (XGBoosting) is a very effective method for data classification [111]. Interestingly, there

has not been a single paper published for breast image classification using the XGBoost algorithm. Along with the boosting method, different bagging method's are available, among them Random Forest (RF) is very popular where a large number of uncorrelated trees are aggregated together for a better prediction. Table 2.4.1 and Table 2.4.1 summarizes a set of papers where a Logic-Based Algorithm has been used for image classification.

Table 2.12: Logic-Based (Part 1)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Beura [112]	(1) Two-dimensional discrete ortho-normal	Mammogram	—	(1) Achieved accuracy and AUC values on MIAS database are 98.3 %, 0.9985.
	S-transform has been used for the feature extraction.			(2) Achieved accuracy and AUC values on DDSM database are 98.8 %, 0.9992.
Diz et al. [113]	(1) GLCM	Mammogram	410	(1) Their achieved accuracy value is 76.60 %
	(2) GLRLM			(2) Mean false positive value is 81.00%.
Zhang et al. [91]	(1) 133 Features (Mass based and Content based)	Mammogram	400	(1) Computer model has been created which is able to find a location that was not detected by trainee.
Ahmad et al. [114]	(1) Nine features selected	Biopsy	700	(1) Achieved Sensitivity, Specificity and Accuracy are 75.00%, 70.00% and 72.00% respectively.
Paul et al. [115]	(1) Harlick Texture feature	Histopathological	50	(1) Their achieved Recall and Precision are 81.13% and 83.50%.
Chen et al. [116]	(1) Dual-tree complex wavelet transform	Mammogram	—	(1) Achieved Received Operating Curve (ROC) 0.764.
	(DT-CWT) has been used for the feature extraction.			
Zhang et al. [117]	(1) Curvelet Transform	Histopathological	50	(1) Random Subspace Ensemble (RSE) utilized.
	(2) GLCM (3) CLBP			(2) Their achieved classification accuracy 95.22% where the previous accuracy on this same database was 93.40%.

Table 2.13: Logic-Based (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Angayarkanni et al. [118]	(1) GLCM.	Mammogram	322	(1) The Achieved Sensitivity and Accuracy are 93.40% and 99.50% respectively
Wang et al. [119]	(1) Horizontal Weighted Sum	Mammogram	322	(1) Surrounding Region Dependence Method (SRDM) utilized for region detection.
	(2) Vertical Weighted Sum			(2) Achieved True Positive Rate 90.00% and False Positive Rate 88.80%.
	(3) Diagonal Weighted Sum			(1) ANOVA method utilized for feature prioritization.
	(4) Grid Weighted Sum.			(2) When they use RF algorithm on Mammogram (DDSM) data-set, obtained accuracy and ROC is 79.00% and 0.89.
Bruno et al. [120]	(1) Curvelet Transform 2. LBP	Mammogram Histopathological	—	(1) Textural features have been extracted from the Regions of Interest (ROIs) using RLTP. (2) They claimed that the RLTP feature provides better performance than the rotation invariant patterns.
Muramatsu et al. [121]	(1) Radial Local Ternary Pattern (RLTP)	Mammogram	376	(1) Chain code utilized for extraction of Region of Interest (ROIs). (2) Rough set method utilized to enhance the ROIs. (3) Their achieved ROC value is 0.947 and obtained Matthews Correlation (MCC) is 0.8652.
Dong et al. [122]	(1) NRL Margin Gradient (2) Gray-level histogram (3) Pixel value fluctuation	Mammogram	—	(1) Local Binary Pattern-Three Orthogonal Projections (LBP-TOP)
Piantadosi et al. [123]	(1) Local Binary Pattern-Three Orthogonal Projections (LBP-TOP)	Mammogram	—	(1) Their achieved Accuracy, Sensitivity and Specificity values are 84.60%, 80.00% 90.90%.

Support Vector Machine (SVM)

SVM machines were proposed by VC (Vepnick-Cherovorenkis). This technique does not require any prior distribution knowledge for the data classification task like Bayesian classification technique. In many practical situations, the distribution of the features is not available. In such cases, SVM can be used to classify the available data into the different classes.

Consider the set of two-dimensional data plotted in Figure 2.17. The symbol "o" represents those data which belong to Class-1 and "□" represents data which belong to Class-2. A hyperplane (P) has been drawn which classifies the data into two classes. Interestingly, there will be "n" hyperplanes available which can separate the data.

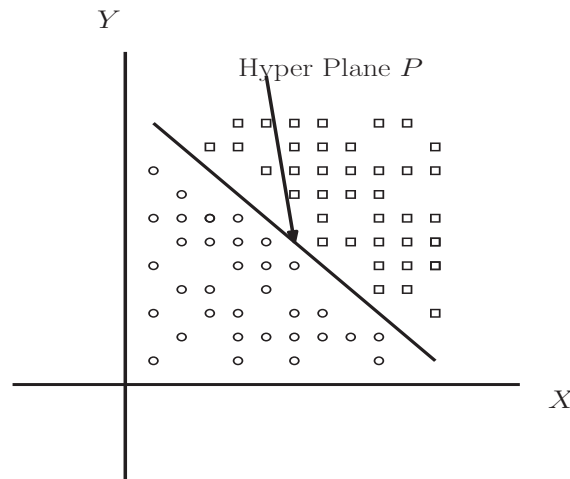


Figure 2.17: SVM finds the hyperplane which separates two classes

Let $\mathbf{X} = \{\mathbf{X}_i\}$ where $\{\mathbf{X}_i \in \mathcal{R}^n\}$ ($i = \{1, 2, 3, \dots, l\}$) is to be classified into two classes $\omega \in \{\omega^1, \omega^2\}$. Suppose that the classes $\{\omega^1\}$ and $\{\omega^2\}$ are recognized as " + 1" and " - 1". Classification of this data can be written

$$\mathcal{C} = \{(\mathbf{X}_1, \omega_1), (\mathbf{X}_2, \omega_2), (\mathbf{X}_3, \omega_3), \dots, (\mathbf{X}_n, \omega_n)\} \quad (2.11)$$

During the learning stage, the SVM finds parameters $\mathbf{W}_i = [W_i^1, W_i^2, \dots, W_i^n]^T$ and b

to produce a decision function $d(\mathbf{X}_i, \mathbf{W}_i, b)$:

$$d(\mathbf{X}_i, \mathbf{W}_i, b) = \mathbf{W}_i^T \mathbf{X}_i + b = \mathbf{W}_i \cdot \mathbf{X}_i + b = \sum_{j=1}^n W_i^j X_i^j + b \quad (2.12)$$

where $\mathbf{W}_i, \mathbf{X}_i \in \mathcal{R}^n$.

As the training data are linearly separable no training data will satisfy the condition

$$d(\mathbf{X}_i, \mathbf{W}_i, b) = 0 \quad (2.13)$$

To control the separability, we consider the following inequalities:

$$d(\mathbf{X}_i, \mathbf{W}_i, b) \geq 1 \quad \text{for} \quad \omega_i = +1 \quad (2.14)$$

$$d(\mathbf{X}_i, \mathbf{W}_i, b) < 1 \quad \text{for} \quad \omega_i = -1 \quad (2.15)$$

Sometime it is very difficult to find the perfect hyperplane which can separate the data, but if we transform the data into a higher dimension the data may be easily separable. To separate this kind of data, a kernel function can be introduced.

Kernel Methods

Assume a transformation ϕ such that it transforms the data-set $\mathbf{X}^1 \in \mathcal{R}^n$ in-to data-set $\mathbf{X}^2 \in \mathcal{R}^m$ where $m > n$. Now train the linear SVM on the data set \mathbf{X}^2 to get a new classifier F_{SVM} .

A Kernel ϕ effectively computes a dot product in a higher-dimensional space \mathcal{R}^m . For $\{\mathbf{x}_i, \mathbf{x}_j\} \in \mathcal{R}^n$, $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i, \mathbf{x}_j) \rangle_m$ is an inner product of \mathcal{R}^m , where $\phi(\mathbf{x})$ transforms \mathbf{x} to \mathcal{R}^m . Consider $\{\mathbf{x}_i, \mathbf{x}_j\} \in \mathcal{R}^n$; then we can define the kernel as follows:

- Radial Basis Function Kernel (rbf): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma | \langle \phi(\mathbf{x}_i - \mathbf{x}_j) \rangle |^2)$
- Polynomial Kernel (polynomial): $K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \phi(\mathbf{x}_i, \mathbf{x}_j) \rangle + r)^d$
- Sigmoid Kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\langle \phi(\mathbf{x}_i, \mathbf{x}_j) \rangle + r)$

- Linear Kernel (linear): $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

The advantage of the kernel method for breast cancer image classification using an SVM was first introduced by El-Naqa et al. [124]. They classify Micro-calcification clusters in mammogram images (76 images were utilized for the experiment where the total number of MCs were 1120). They utilized the SVM method along with the Gaussian Kernel as well as the polynomial kernel. In 2003, R.Chang et al. classified a set of sonography images using SVM techniques where they consider that the image is surrounded by pickle noise [125], where the data-base contains 250 images. Their achieved accuracy was 93.20%. A total of thirteen features, including Shape, Law's and gradient features, were utilized along with SVM and a Gaussian kernel for the mammogram image classification. They performed their operation on 193 mammogram images and achieved 83.70% sensitivity and 30.20% false positive rate [126]. SVM has been combined with the NN method by Sing et al. for ultrasound breast image classification where the database contained total 178 images. They performed a hybrid feature selection method to select the best features [127].

A breast ultrasound image is always very complex in nature. The Multiple Instance Learning (MIL) algorithm has been first used along with SVM for the breast image classification by [128], and their obtained accuracy was 91.07 %. The Concentric Circle BOW Feature-Extraction method was utilized to extract the features and later the SVM method was used for breast image classification [129]. Their achieved accuracy is 88.33% when the dimension of the features was 1000. A Bag-of-Features has been extracted from Histopathological images (using SIFT and DCT), and using SVM for classification by Maa et al. [130]. The experiment is performed on a database which contains 361 images, where 119 images are normal, 102 images are ductal carcinoma in situ and the rest of the images are invasive carcinoma. Their experiment achieved 100.00% classification accuracy for ductal carcinoma in situ, 98.88% classification accuracy for for invasive carcinoma, and

100.00% classification accuracy for normal image classification. A mammogram (DDSM) image database has been classified by Hiba et al. [131] by SVM along with the Bag of Feature method. Firstly the authors extract LBP and quantize to the binary pattern information for Feature-Extraction. Their obtained accuracy was 91.25%.

Along with the above-mentioned work different breast image databases have been analyzed and classified using SVM. We have summarized some of the work related to SVM in Tables 2.14, 2.15 and 2.16.

Table 2.14: SVM for Breast-Image Classification (Part 1)

Reference	Descriptor	Image Type	No. of Images	Key Findings
B Malik et al. [132]	(1) Speed of sound	QTUS	—	(1) Glands, Fat, Skin, and Connective tissue have been classified.
	(2) Attenuation image vector			(2) Both Linear and nonlinear SVM classifier have been utilized.
	(3) Reflection image vector			(3) Their experiment obtained 85.20% accuracy.
Chang et al. [133]	(1) Textural Features such as	Ultrasound	250	(1) Benign and Malignant images have been classified.
	i. Auto Correlation Coefficient			(2) Accuracy, Sensitivity, Specificity, Positive predictive values, Negative predictive value are 85.60 %, 95.45%, 77.86, 77.21 %
	ii. Auto Covariance Coefficient			and 95.61 % respectively.
Akbay et al. [134]	(1) 52 features have been extracted.	Mammogram	—	(1) Micro-calcification (MC) Classification accuracy 94.00%
	(1) Relative Signal Intensities			
	(2) Derivative of Signal Intensities			
Leyman et al. [135]	(3) Relative Signal Intensities and their Derivatives in one Vector	MRI	76	(1) Benign and Malignant lesions are investigated.
	(4) i) Maximum of Signal intensity enhancement ii) time of maximum enhancement iii) time of maximum washout			(2) Linear kernel, a polynomial kernel and a radial basis function kernel utilized along with the SVM method for the breast image classification.
Martins et al. [136]	(1) Ripley's K Function	Mammogram	390	(1) Benign and Malignant image classification.
				(2) The achieved Accuracy, Sensitivity and Specificity are 94.94%, 92.86% and 93.33% respectively.

Table 2.15: SVM for Breast-Image Classification (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Zhang et al. [137]	(1) Fractional Fourier transform	Mammogram	200	(1) They selected ROI for avoiding redundant complexity.
	information utilized as features			(2) When SVM and Principal Component Analysis were used together the achieved Accuracy, Sensitivity and Specificity are 92.16 ± 3.60 % , 92.10 ± 2.75 % and 92.22 ± 4.16 % respectively.
F.Shirazi et al. [138]	(1) GLCM	Ultrasound	322	(1) ROI extracted for reducing redundant complexity.
				(2) SVM and Mixed Gravitational Search Algorithm (MGSA) used together for feature reduction. (3) The achieved accuracy 86.00%, however SVM with MGSA method achieved 93.10% accuracy.
Sewak et al. [139]	(1) radius, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension and texture of nuclei calculated.	Biopsies	569	(1) Achieved Accuracy, Sensitivity and Specificity are 99.29%, 100.00% and 98.11% respectively.
Dheba et al. [140]	(1) The Laws Texture features utilized.	Mammogram	322	(1) The achieved accuracy is 86.10%.

Table 2.16: SVM for Breast-Image Classification (Part 3)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Taheri et al. [141]	(1) Intensity information	Mammogram	600	(1) Classified images into normal and abnormal images.
	(2) Value of detected corner			(2) Removing unwanted objects from the images for reducing the redundancy and computational complexity.
	3. Energy			(3) Achieved Precision and Recall rates are 96.80% and 92.5% respectively.
Tan et al. [142]	(1) Shape, Fat, Presence of Calcification Texture, Spiculation,	Mammogram	1200	(1) Features have been selected from the region of interest.
	Contrast, Isodensity type features selected			(2) They utilized the radial basis function (RBF) for their analysis.
	(2) Total number of features 181			(3) The Sequential Forward Floating Selection (SFFS) method utilized for the feature selection. (4) The area under the receiver operating characteristic curve was (AUC)= 0.805± 0.012.
Kavitha et al. [143]	(1) Histogram of the intensity has been used as a statistical feature.	Mammogram	322	(1) When using SVM with the linear kernel the obtained Accuracy, Sensitivity and Specificity are 98%, 100% and 96% respectively.
	(2) 2-D Gabor filter utilized for the textural feature extraction			(2) When using Weighted Feature SVM with weights the obtained Accuracy, Sensitivity and Specificity are 90%, 100% and 75% respectively.
	(3) Clinical features extracted from the data-base directly.			

Bayesian

A Bayesian classifier is a statistical method based on Bayes theorem. This method does not follow any explicit decision rule, however it depends on estimating probabilities. The Naive Bayes method can be considered one of the earlier Bayesian learning algorithms.

The Naive Bayes (NB) method works on the basis of the Bayes formula, where each of the features is considered statistically independent. Consider a data-set with m samples, each sample containing a feature vector \mathbf{x}^k with n features [144] and belonging to a particular class c_k . According to the NB formula, the probability of the particular class c_k with the conditional vector \mathbf{x}^k is represented as

$$P(c_k|\mathbf{x}^k) = \frac{P(\mathbf{x}^k|c_k)P(c_k)}{P(\mathbf{x}^k)} \quad (2.16)$$

Applying the chain rule

$$P(\mathbf{x}_1^k, \mathbf{x}_2^k, \mathbf{x}_3^k, \dots, \mathbf{x}_n^k|c_k) = \prod_{i=1}^n P(\mathbf{x}_i^k|c_k) \quad (2.17)$$

The NB theorem considers all the features independently which can be represented as

$$\bar{c} = \arg \max_{k \in 1, \dots, m} P(c_k) \prod_{i=1}^n P(\mathbf{x}_i^k|c_k) \quad (2.18)$$

The NB method is very easy to construct and very fast to predict the data. This method can also be utilize the kernel method. However, for a large data-set and continuous data, this method has very poor performance. NB can be classified into the following subclasses:

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bernoulli Naive Bayes

One of the constraints of the NB classifier is that it considers that all the features are conditionally independent. A Bayesian Network is another Bayesian classifier which can overcome this constraint [145], [146]. The literature shows that, the Bayesian classifier method is not utilized much for breast image classification. In 2003 S. Butler et al. used

NB classifier for X-ray breast image classification [147]. They extracted features from the low-level pixels. For all feature combinations they obtained more than 90.00% accuracy. Bayesian structural learning has been utilized for a breast lesion classifier by Fishcer et al. [148]. D. Soria et al. [149] classify a breast cancer data-set utilizing C4.5, multilayered perceptron and the NB algorithm using WEKA software [150]. They conclude that the NB method gives better performance than the other two methods in that particular case. They also compared their results with the Bayes Classifier output. Some other research on the Bayes classifier and breast image classification has been summarized in Table 2.17 and 2.18.

Table 2.17: Bayesian Classifier (Part 1)

Reference	Descriptor	Image Type	No. of Image.	Key Finding.
Kendall et al. [151]	(1) Features extracted using DCT method.	Mammogram		(1) Bayesian classifier obtained 100.00% sensitivity with 64.00% specificity.
Oleksyuk et al. [152]		—	—	(1) Bayesian method obtained 86.00% with 80.00% specificity.
Claridge et al. [153]	(1) Statistical and LBP fetures extracted.	Mammogram	322/410	(1) Bayesian method obtained 67.07 ± 0.73 % and 67.61 ± 0.83 % accuracy on MIAS and Inbreast image data-sets (using statistical features). (1) Bayesian method obtained 62.86 ± 0.70 % and 51.99 ± 1.28 % Accuracy on MIAS and Inbreast image data-sets (using LBP).
Raghavendra et al. [154]	(1) Gabor wavelet transform utilized for feature extraction.	Mammogram	690	(1) Locality Sensitive Discriminant Analysis (LSDA) for the data reduction. (2) NB obtained 84.34% Accuracy, 83.69% Sensitivity with 90.86% Specificity.
Perez et al. [155]	(1) 23 features utilize.	Mammogram	—	(1) UFilter feature selection methods utilized and its efficiency verified by Wilcoxon statistical test.
Rashmi et al. [156]	(1) 10 features utilized.	—	—	(1) Benign and Malignant tumors have been classified.
Gatuha et al. [157]	(1) 10 features utilized	—	—	(1) They built an android based Benign and Malignant tumor classifier. (2) Their obtained accuracy is 96.4%

Table 2.18: Bayesian Classifier (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Bemendorf et al. [158]	(1) BI-RADS features utilized.	—	2766	(1) For the training data the AUC value is 0.959 for the inclusive model, whereas AUC value is 0.910 for the descriptor model.
Rodriguez et al. [159]	(1) Eight Image feature nodes utilized.	—	—	(1) NB model obtained 79.00% Accuracy, 80.00% Sensitivity.
Nugroho et al. [160]	(1) Eight Image feature nodes utilized.	Mammogram	—	(1) Naive Bayes model along with SMO; obtained ROC value is 0.903. (2) Bayesian Network model along with SMO; obtained accuracy was 83.68%.
Rodriguez et al. [161]	(1) Eight Image features have utilized	—	231	(1) Bayesian Network model obtained 82.00% accuracy, 80.00% sensitivity and 83.00% Specificity when they utilized only three features.
Shivakumari et al. [162]		—	231	1. Analyze the Ljubljana breast image data set. (2) NB algorithm along with feature ranking techniques; the best achieved Accuracy was 81.46%.
Rodriguez et al. [163]	(1) Seven different clinical features extracted	Mammogram	690	(1) Obtained Accuracy, Sensitivity and Specificity are 82.00%, 80.00% and 83.00% respectively.

2.4.2 Performance Based on Un-supervised Learning

This learning algorithm doesn't require any prior knowledge about the target. The main goal of the un-supervised learning is to find the hidden structure and relations between the different data [164] and distribute the data into different clusters. Basically a clustering is a statistical process where a set of data points is partitioned into a set of groups, known as a cluster. The K-means algorithm is a clustering algorithm proposed by [165]. Interestingly, unsupervised learning can be utilized as pre processing step too.

- In the K-means algorithm, firstly assign K centroid points. Suppose that, we have n feature points x_i where $i \in \{1, \dots, n\}$. The objective of the K-means algorithm is to find positions μ_i , where $i \in 1, \dots, K$ that minimize the total variance of the clusters by solving

$$\arg \min_{x \in c_i} \sum_{i=1}^K \sum_{x \in c_i} d(x, \mu_i) = \arg \min_{x \in c_i} \sum_{i=1}^K \sum_{x \in c_i} \|x - \mu_i\|^2 \quad (2.19)$$

- Self-Organizing Map (SOM)

SOM is another popular unsupervised classifier, proposed by Kohonen et al. [166], [167], [168]. The main idea of the SOM method is to reduce the dimension of the data and represent those dimensionally reduced data by a map architecture, which provides more visual information.

- Fuzzy C-means clustering (FCM)

The FCM algorithm cluster data based on the value of a membership function, is proposed by [169] and improved by Bezdek [170].

The history of using unsupervised learning for breast image classification is a long one . In 2000, Cahoon et al. [171] classified mammogram breast images (DDSM data-base) in an un-supervised manner, utilizing the K-NN clustering and Fuzzy C-Means (FCM) methods.

Chen et al. classified a set of breast images into benign and malignant classes [172]. They utilized a SOM procedure to perform this classification operation. They collected 24 autocorrelation textural features and used a 10 Fold validation method. Markey et al. utilized the SOM method for BIRADS image classification of 4435 sample [173]. Table 2.19 and 2.20 summarises the breast image classification performance based on K-means algorithm and SOM method.

Table 2.19: K-means Cluster Algorithm and Self-Organising Map for Breast-Image Classification (Part 1)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Moftah et al. [174]	(1) Intensity distribution used as feature.	MRI	—	(1) Three types of evaluation measures performed: a) Accuracy, b) Feature based, c) Shape based measure. (2) This can classify the data as well as identify the target. (3) The obtained best accuracy of the segmented ROI is 90.83%
Lee et al. [175]	(1) 1734 Signal Patterns	MRI	322	(1) Available signal patterns have been classified into 10 classes.
Dalmiya et al. [176]	(1) Discrete Wavelet Transform.	Mammogram	—	(1) Cancer tumor masses have been segmented.
Elmoufidi et al. [177]	(1) Local Binary Pattern.	Mammogram	322	(1) Image enhancing. (2) Generation of number of clusters (3) Detection of regions of interest. (4) Mean detection of regions of interest is 85.00%.
Samundeeswari et al. [178]		Ultra Sound	—	(1) Utilizing Ant Colony and Regularization parameters. (2) This method obtained 96.00% similarity between segmented and reference tumors.
Rezaee [179]	(1) Discrete Wavelet Transform	Mammogram	120	(1) Early detection of tumors from the breast image. (2) Tumor detection Accuracy 92.32%, sensitivity 90.24%.
Chandra et al. [180]	(1) Gray Intensity values	Mammogram	—	(1) Mammogram image has been clustered using SOM along with the Quadratic Neural Network.

Table 2.20: K-means Cluster Algorithm and Self-Organising Map for Breast-Image Classification (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Lashkari et al. [181]		Thermogram	23	(1) Both FCM method and Adaboost method utilized separately to classify images. (2) For the classification purposes selected 23 features and also select the best features using feature selection algorithm. When they used the FCM method, the obtained mean accuracy was 75.00% whereas the Adaboost method accuracy was 88.00%.
Nattkemper et al. [182]		MRI	—	(1) K-means algorithm as well as SM method utilized.
Slazar-Licea et al. [183]. Marcomini et al. [184]	(1) 24 morphological features	...	—	(1) Fuzzy c-means algorithm used. (1) Minimizing noise using Wiener filter, equalized and median filter
Chen et al. [172]	(1) 24 autocorrelation texture features	Ultra Sound	243	(2) Obtained Sensitivity 100 % and Specificity 78.00%. (1) Obtained ROC area 0.9357 ± 0.0152 . accuracy 85.60%, Specificity 70.80%.
Iskan et al. [185]	(1) Two-dimensional discrete cosine transform (2) 2D continuous wavelet transform	Ultra Sound	—	(1) Automated threshold scheme introduce to increase the robustness of the SOM algorithm.

2.4.3 Performance Based on Semi-supervised Learning

The working principle of semi-supervised learning lies in between supervised and unsupervised learning. For the semi-supervised learning a few input data have an associated target and large amounts of data are not labeled [186]. It is always very difficult to collect the labeled data. Few data such as speech or information scratched from the web are difficult to label. To classify this kind of data semi-supervised learning is very efficient. However lately this method has been utilized for the breast image classification too. Semi-supervised learning can be classified as

- Graph Based (GB)
- Semi-Supervised Support Vector Machine
- Human Semi-supervised Learning

To the best of our knowledge, Li et al. has been utilized GB semi-supervised learning for biomedical image classification [187]. The kernel trick is applied along with the semi-supervised learning method for breast image classification by Li et al. [188]. They performed their experiments on the Wisconsin Prognostic Breast Cancer (WPBC) dataset for the breast image classification. Ngadi et al. utilized both the SKDA (Supervised Kernel-based Deterministic Annealing) and NSVC methods for mammographic image classification [189]. They performed their experiments on 961 images, where 53.60% of the images were benign and the rest of the images are malignant. Among the other utilized features they utilized BI-RADS descriptors as features. When they utilized the NSVC method they also utilized RBF, Polynomial and linear kernel. They found that the best accuracy of 99.27% was achieved when they utilized linear kernels. Few studies have performed the breast image classification by semi-supervised learning, as summarized in Table 2.21 and 2.22 .

Table 2.21: Semi-Supervised Algorithm for Breast-Image Classification (Part 1)

Reference	Descriptor	Image Type	No of Image	Key Finding
Cordeiro et.al. [190]	(1) Zernike moments has been used for the feature extraction.	—	685	(1) Semi-supervised Fuzzy GrowCut algorithm utilized. (2) For the fatty-tissue classification this method achieved 91.28% accuracy.
Cordeiro et al. [191]	—	Mammogram	322	(1) Semi-supervised Fuzzy GrowCut as well as the Fuzzy GrowCut algorithm utilized for tumors, region segmentation.
Nawel et al. [192]	—	—	—	(1) Semi-supervised Support Vector Machine (S3VM) utilized. (2) This experiment shows impressive results on the DDSM data-base.
Zemmal et al. [193]	—	DDSM	—	(1) Transductive Semi-supervised learning technique using (TSVM) utilized for classification along with different features.
Zemmal et al. [194]	—	—	200	(1) Semi-supervised Support Vector Machine (S3VM) utilized with various kernels.
Nawel et al. [195]	(1) GLCM 2. Hu moments (3) Central Moments	Mammogram	—	(1) Transductive Semi-supervised learning technique used for image classification. (2) This experiment shows impressive results on DDSM data-base.
Peikari et al. [196]	(1) Mean, Mode, Standard Deviation, Media, Skewness, Kurtosis	Histopathological	322	(1) The Ordering Points to Identify the Clustering Structure (OPTICS) method utilized for image classification [197].

Table 2.22: Semi-Supervised Algorithm for Breast-Image Classification (Part 2)

Reference	Descriptor	Image Type	No. of Images	Key Findings
Zhu et al. [198]	(1) Relative local intensity	Ultra Sound	144	(1) One important micro-environment inside the tumor is vasculature, which has been classified in this paper.
	(2) Shape irregularity			
	(3) Orientation Consistency			
Liu et al. [199]	—	Ultra Sound	—	(1) Iterated Laplacian regularization based semi-supervised algorithm for robust feature selection (Iter-LR-CRFS) utilized.
				(2) The archived accuracy and sensitivity are $89.0\pm3.6\%$ and $91.0\pm5.2\%$

2.5 Conclusion

Breast-cancer is a serious threat to women throughout the world and is responsible for increasing the female mortality rate. The improvement of the current situation with breast cancer is a big concern and can be achieved by proper investigation, diagnosis and appropriate patient and clinical management. Identification of breast cancer in the earlier stages and a regular check of the cancer can save many lives. The status of cancer changes with time, as the appearance, distribution and structural geometry of the cells is changing on a particular time basis because of the chemical changes which are always going on inside the cell. The changing structure of cells can be detected by analysing biomedical images which can be obtained by Mammogram, MRI etc. techniques. However these images are complex in nature and require expert knowledge to perfectly analyze for malignancy. Due to the non-trivial nature of the images the physician sometimes makes a decision which might contradict others. However computer-aided-diagnosis techniques emphasising on the machine learning can glean a significant amount of information from the images and provide a decision based on the gained information, such as cancer identification, by classifying the images.

The contribution of Machine-Learning techniques to image classification is a long story. Using some advanced engineering techniques with some modifications, the existing Machine-Learning based image classification techniques have been used for biomedical image classification, specially for breast-image classification and segmentation. A few branches of the Machine-Learning based image classifier are available such as Deep Neural Network, Logic based, SVM etc. Except for deep learning, a machine learning-based classifier largely depends on handcrafted feature-extraction techniques such as statistical, structural information etc., that depend on various mathematical formulations and theories where they gain object-specific information. They are further utilized as an input for an image classifier such as SVM, Logic based etc, for the image classification.

This investigation finds that, most of the conventional classifiers depends on prerequisite local Feature-Extraction. The nature of cancer is always changing, so the dependencies on a set of local features will not provide good results on a new dataset. However the state-of-the art Deep Neural Networks, specially CNN, have recently advanced biomedical image-classification due to the global-feature extraction capabilities. As the core of the CNN model is the kernel, which gives this model the luxury of working with the global features. These globally extracted features allow the CNN model to extract more hidden structure from the images. This allows some exceptional results for Breast-Cancer image-classification. As the CNN model is based on the global features, this kind of classifier model should be easy to adapt to a new dataset.

This paper also finds that the malignancy information is concentrated in the particular area defined as ROI. Utilizing only the ROI portions, and information gathered from the segmented part of the data can improve the performance substantially. The recent development of the Deep Neural Network can also be utilized for finding the ROI and segmenting the data, which can be further utilized for the image-classification.

For Breast-Cancer patient care, the Machine-Learning techniques and tools have been a tremendous success so far, and this success has gained an extra impetus with the involvement of deep learning-techniques. However the main difficulty of handling the current deep-learning based Machine-Learning classifier is its computational complexity, which is much higher than for the traditional method. The current research is focused on the development of the light DNN model so that both the computational and timing complexities can be reduced. Another difficulty of using the DNN based cancer image-classifier is that it requires a large amount of training data. However the reinforcement of learning techniques and data augmentation has been largely adapted with the current CNN model, which can provide reliable outcomes. Our research finds that the current trend of Machine-Learning is largely towards deep-learning techniques. Among a few other impli-

cations, the appropriate tools for designing the overall deep-learning model was the initial obligation for utilizing deep-learning based Machine-Learning techniques. However some reliable software has been introduced which can be utilized for breast-image classification. Initially it was difficult to implement a DNN-based architecture in simpler devices, however due to cloud-computer based artificial-intelligence techniques this issue has been overcome and DNN has already been integrated with electronic devices such as Mobile phones. In future combining the DNN network with the other learning techniques can provide more-positive predictions about breast cancer.

Due to the tremendous concern about breast cancer, many research contributions have been published so far. It is quite difficult to summarize all the research work related to Breast-Cancer image-classification based on Machine-Learning techniques in a single research article. However this chapter has attempted to provide a holistic approach to the Breast-Cancer image- classification procedure which summarises the available breast dataset, generalized image-classification techniques, Feature-Extraction and reduction techniques, performance measuring criteria and state-of-the-art findings.

In a nutshell, the involvement of machine learning for breast-image classification allows doctors and physicians to take a second opinion, and it provides satisfaction to and raises the confidence level of the patient. There is also a scarcity of expert people who can provide the appropriate opinion about the disease. Sometimes the patient might need to spend a long time waiting due to the lack of expert people. In this particular scenario the machine learning based diagnostic system can help the patient to receive the timely feedback about the disease which can improve the patient-management scenario.

Chapter 3

Histopathological Breast-Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering

3.1 Abstract

Breast-Cancer is a serious threat and one of the largest causes of death of women throughout the world. The identification of cancer largely depends on digital biomedical photography analysis such as Histopathological images by doctors and physicians. Analysing Histopathological images is a non-trivial task, and decisions from investigation of these kinds of images always require specialised knowledge. However, Computer Aided Diagnosis (CAD) techniques can help the doctor to make more reliable decisions. The state-of-the-art Deep Neural Network (DNN) has been recently introduced for biomedical image analysis. Normally each image contains structural and statistical information. This

Published as: AA. Nahid, M. A. Mehrabi and Y. Kong, "Histopathological Breast-Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering", *BioMed Research International*, Hindawi, pp. 1-20, 2018.

chapter classifies a set of biomedical breast-cancer images (BreakHis dataset) using novel DNN techniques guided by structural and statistical information derived from the images. Specifically a Convolutional Neural Network (CNN), a Long-Short-Term-Memory (LSTM) and a combination of CNN and LSTM are proposed for Breast-Cancer image classification. Softmax and Support Vector Machine (SVM) layers have been used for the decision-making stage after extracting features utilising the proposed novel DNN models. In this experiment the best Accuracy value of 91.00% is achieved on the $200\times$ dataset, the best Precision value 96.00% is achieved on the $40\times$ dataset, and the best F-Measure value is achieved on both the $40\times$ and $100\times$ datasets.

3.2 Introduction

The unwanted growth of cells causes cancer which is a serious threat to humans. Statistics show that millions of people all over the world suffer various cancer diseases. As an example Table 3.1 summarises the statistics concerning the recent cancer situation in Australia. These statistics reveal the number of newly cancer-affected people diagnosed in Australia and also the number of people who died in 2017 in Australia. These statistics also divulge that the number of females affected and the number of females dying due to breast cancer are more than for males. This indicates that females are more vulnerable to Breast-Cancer (BC) than males. Although these statistics are for Australia they might be representative of what is happening throughout the world.

Table 3.1: Cancer Statistics for Australia 2017 [2]

	Female	Male	Total
Estimated number of new diagnoses (all cancers)	62005	72169	134174
Estimated number of deaths	20677	27076	47753
Estimated new case of diagnosis (Breast Cancer)	17586	144	17730
Deaths due to Breast Cancer	3087	57	3114

Proper BC diagnosis can save thousands of women's lives, and proper diagnosis largely depends on identification of the cancer. Finding BC largely depends on capturing a photograph of the cancer-affected area which gives information about the current situation of the cancer. A few biomedical imaging techniques have been utilised, some of which are non-invasive such as Ultra-Sound imaging, X-Ray imaging, Computer Aided Tomography (CAT) imaging. Other imaging techniques are invasive such as Histopathological images. Investigation of these kinds of images is always very challenging, specially in the case of Histopathological imaging due to its complex nature. Histopathological image analysis is non-trivial, and the investigation of this kind of image always produces some contradictory decisions by doctors. Since doctors and physicians are human, it is natural that errors will occur.

A Computer Aided Diagnosis (CAD) system provides doctors and physicians with valuable information, for example classification of the disease. Different research groups investigate opportunities to improve the CAD systems' performance. Some advanced engineering techniques have been utilised to take a general image classifier and adjust it as a biomedical image classifier, such as a breast-image classifier. The state-of-the-art Deep Neural Network (DNN) techniques have been adapted for a BC image classifier to provide reliable solutions to patients and their doctors.

The basic working principle of DNN lies on the basic Neural Network (NN). Rosenblatt in 1957 [45] for the very first time introduced the NN concept, which provides decisions based on a threshold. Using some advanced engineering, a very light Convolutional Neural Network (CNN) model has been proposed by K. Fukushima [79], referred to as "Neocognitron". The main interest of this project is to find stimulus patterns, where they can tolerate a limited amount of shifting variance. This "Neocognitron" model served as the first CNN model for biomedical signal analysis [79]. Specifically a CNN model has been for the first time introduced for breast-image classification by Y. Wu et al. [80] where

they performed their experiments on a set of mammogram images. The utilisation of the CNN model for breast-image classification has been limited due to its computational complexity, until A. Krizhevsky et al. [82] proposed their model known as AlexNet. This AlexNet model has brought about a revolutionary change in the image-analysis field, specially image classification. Taking this model as a reference, a few other models have been adjusted such as ResNet [84], Inception [85], Inception-V4, Inception-ResNet [86], etc., for biomedical image classification. M. A. Jaffar et al. classified the mammogram-image (MIAS-mini, DDSM) dataset using the CNN model, and obtained 93.35% Accuracy and 93.00% Area Under Curve (AUC) [200]. Y. Qiu et al. [101] utilised a CNN for mammogram image classification and where they utilised 2, 5 and 10 feature maps and obtained an average Accuracy of 71.40%. M. G. Ertosum et al. [201] employed the CNN method for automated positioning of the masses as well as breast-image classification and obtained 85.00% Accuracy. Y. Qui et al. [202] classified a set of mammogram images into benign and malignant classes, where they utilized a total of 560 Regions of Interest (ROI). Z. Jiao et al. [97] characterised a set of mammogram images into benign and malignant images and obtained 96.70% Accuracy. A set of mammogram images has been classified by B. Sahiner et al., and their achieved ROC score is 0.87 [81]. M. M. Jadoon et al. classified a set of mammogram breast images into normal, benign and malignant classes utilising a CNN model.

As with mammogram images, Histopathological breast images have been classified by different research groups. Referring to the most recent, Y. Zheng et al. classify a set of Histopathological images into benign and malignant classes by locating the nucleus from the images using the blob detection method [203]. T. Araujo et al. classify a set of Histopathological images utilising CNN into four classes (normal tissue, benign tissue, in situ carcinoma and invasive carcinoma) and two classes (carcinoma and non-carcinoma). For the four-class classification they obtained 77.80% Accuracy, and when they performed

the two-class classification they obtained 83.3% Accuracy [204]. F. A. Spanhol et al. utilised a CNN model and classified Histopathological images from the BreakHis dataset containing four sets of images based on the magnification factor. They obtained a best image classification Accuracy of $85.6 \pm 4.88\%$ when they utilised the $40\times$ magnification dataset [205].

Images normally preserved a local as well as a hidden pattern which represent similar information. Histopathological images represent different observations of biopsy situation. The biopsy images which belong to the same groups normally preserve similar kinds of knowledge. Unsupervised learning can detect this kind of hidden pattern. The main contribution of this chapter is to classify a set of biomedical breast cancer images using proposed novel DNN models guided by an unsupervised clustering method. Three novel DNN architectures are proposed based on a Convolutional Neural Network (CNN), a Long-Short-Term-Memory (LSTM), and a combination of the CNN and LSTM models. After the DNN model extracts the local and global features from the images the final classification decision is made by the classifier layer. As the classifier layer, this chapter has utilised both the Softmax layer and a Support Vector Machine (SVM). Figure 3.1 demonstrates the overall image classifier model which has been utilised in this experiment.

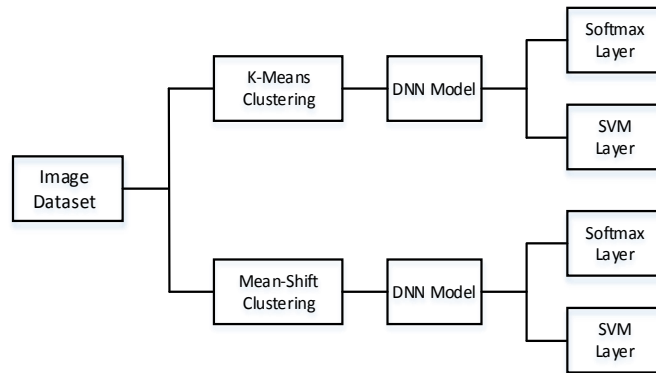


Figure 3.1: Overall image-classifier model for benign and malignant image classification

The remainder of this chapter is organized as follows. Section 3.3 describes the feature

partitioning method based on clustering techniques. Section 3.4 describes DNN models and this is followed by Section 3.5 which describes our proposed novel model based on the DNN method for the breast image classification. Section 3.6 describes and makes a detailed analysis of the results. Section 3.7 compares our findings with existing state-of-the-art findings; and lastly Section 3.8 concludes the chapter.

3.3 Feature partitioning

Images naturally contain significant amounts of statistical and geometrical information. Representation of this kind of structural learning is a prior step for many data analysis procedures such as image classification. One of the techniques of finding the structural information is clustering the data in an unsupervised manner. Clustering allows the same kind of vector to be partitioned into the region. The clustering method partitions data of a similar nature and information in such a way that the partition between the grouped data is maximised. A few clustering methods are available. To find the hidden structure of the data, in this chapter we use the K-Means and Mean-Shift clustering algorithm approaches, which have been explained as follows:

- The K-Means (KM) algorithm is easy to implement, is less computationally complex and can be calculated as follows:

Algorithm 1 K-Means Algorithm [206]

- 1: Consider a set of data points $x_n \in R^D$ where $n \in \{1, 2, \dots, N\}$
- 2: Consider the cluster set $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_c\}$; here $|\mathcal{C}|$ represents the number of cluster, and their corresponding centroid point is $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_c\}$
- 3: Any data point x_i assigned to a particular cluster based on \mathcal{C}_c is given by

$$\mathcal{C}_n = \arg \min_{\mathcal{E}_n \in \mathcal{E}} \text{dist}(\mathcal{E}_n, x_i)^2 \quad (3.1)$$

- 4: Calculate the new centroid:

$$\mathcal{E}_n = \frac{1}{|\mathcal{C}_n|} \sum_{x_i \in \mathcal{C}_n} x_i \quad (3.2)$$

- 5: If no new data point is found stop the search.
-

- The Mean-Shift (MS) algorithm by nature is nonparametric and does not have any assumption about the number of clusters. The MS algorithm can be described as follows:

Algorithm 2 Mean-Shift Algorithm [207]

- 1: Assume a set of data points $x_n \in R^D$ where $n \in \{1, 2, \dots, N\}$. Define a neighbour determining function \mathcal{N}_x , which actually represents a window.
- 2: For $n=1:n$
- 3: Find neighbouring points, of x_i using the function \mathcal{N}_x
- 4: Calculate the MS value

$$\mathcal{MS} = \frac{\sum_{x \in \mathcal{N}_x} \mathcal{K}(x_i - x_n) \times x_i}{\sum_{x \in \mathcal{N}_x} \mathcal{K}(x_i - x_n)} \quad (3.3)$$

- 5: $x_i \leftarrow \mathcal{MS}$
 - 6: Run the algorithm until any new \mathcal{MS} is found.
-

Figure 3.2 shows a benign and a malignant image and their clustering images.

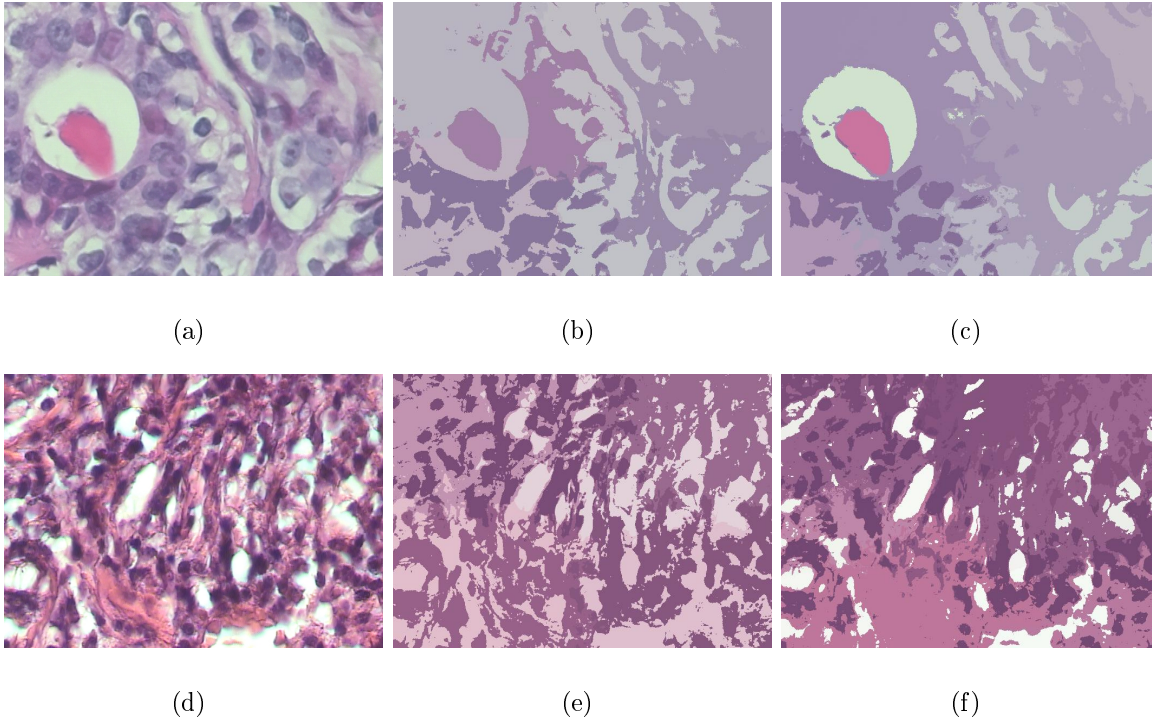


Figure 3.2: Figures a, b, c represent an original benign image, the KM cluster-transformed image, and the MS cluster-transformed image, respectively. Figures d, e, f represent an original malignant image, the KM cluster-transformed image and the MS cluster-transformed image, respectively.

3.4 Deep Neural Network

A Deep Neural Network is a state-of-the art technique for data analysis and classification. A few different DNN models are available, among them the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). They have made some revolutionary improvements in the data analysis field. The following subsection will present the working principle of CNN and RNN (specially on the Long-Short-Term-Memory algorithm), and the working mechanism of the combination of the CNN and LSTM methods.

3.4.1 Convolutional Neural Network

A CNN model is an advanced engineering version of a conventional neural network where the convolution operation has been introduced, which allows the network to extract local as well as global features from the data, enhancing the decision-making procedure of the network. To perfectly control the work flow of a CNN network, along with a convolutional layer, a few intermediate layers have been introduced. These are explained in more detail below.

Convolutional layer

A convolutional layer has been considered to be the main strength or key mechanism for the overall CNN model. In the convolutional layer the value of each position (m_1, m_2) of the input data $I_{m_1 \times m_2}$ has been convolved with the kernel $K_{k_1 \times k_2}$ to produce the feature map. The convolutional output of layer l and feature t for a particular data point (m_1, m_2) of the input data $I_{m_1 \times m_2}$ can be written as

$$I_{m_1, m_2} \star K_{k_1 \times k_2} = \sum_{i=\frac{-k_1+1}{2}}^{\frac{k_1-1}{2}} \sum_{j=\frac{-k_2+1}{2}}^{\frac{k_2-1}{2}} I_{m_1-i, m_2-j} \star K_{i \times j}. \quad (3.4)$$

After adding the bias term $\mathcal{B}^{(l,t)}$ the previous equation will be

$$\mathcal{F}^{(l,t)} = (I_{m_1, m_2} \star K_{k_1 \times k_2}) + \mathcal{B}^{(l,t)}. \quad (3.5)$$

Each of the neurons produces a linear output. When the output of a neuron is fed to another neuron, it eventually produces another linear output. To overcome this issue nonlinear activation functions such as

- Sigmoid
- TanH
- ReLU
- Leaky-ReLU

have been introduced.

Figure 3.3 (a) represents the Sigmoid function characteristic which follows the equation

$$\sigma(x) = \frac{1}{(1 + e^{-x})} \quad (3.6)$$

Interestingly this method suffers due to vanishing gradient problems and having large computational complexity. Another nonlinear activation function is TanH which is basically a scaled version of the $\sigma(x)$ operator such as

$$\tanh(x) = 2 \times \sigma(x) - 1. \quad (3.7)$$

which can avoid the vanishing gradient problem and its characteristics are presented in Figure 3.3 (b). The most popular nonlinear operator is Rectified Linear Unit (ReLU), which filters out all the negative information (like Figure 3.3 (c)) and is represented by

$$\text{ReLU}(x) = \max(0, x). \quad (3.8)$$

Figure 3.3 (d) shows the Leaky-ReLU rectifiers's characteristics, which is a modification of ReLU:

$$\text{Leaky-ReLU}(x) = \sigma(x) + \beta \text{ReLU}(x) \quad (3.9)$$

where β is a predetermined parameter.

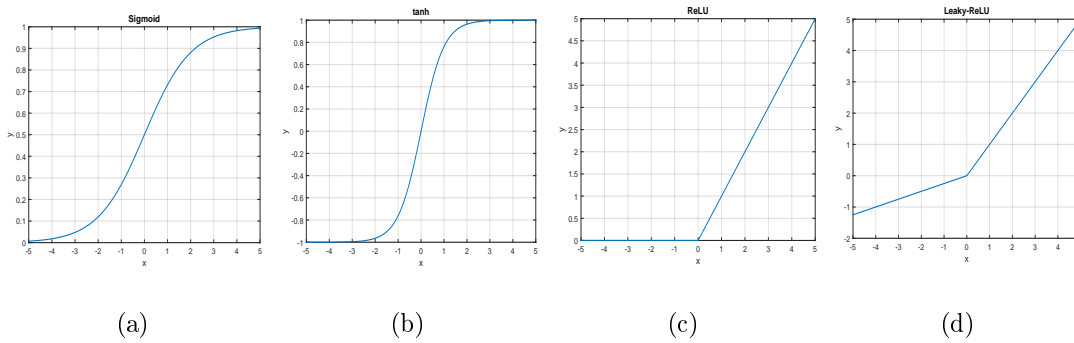


Figure 3.3: Sigmoid, TanH, ReLU and Leaky ReLU

The main ingredient of the convolutional layer is the kernel, which scans through all the input data and tries to extract the global features. The number of steps a kernel

moves each time is known as the stride. The border row and column positions might not be convolved perfectly if we select imperfect stride steps and size. To perfectly conduct the convolution operation at the border, few extra rows and columns (with all zeros) are added, which is known as zero padding.

The convolutional model produces a significant amount of feature information. As the model structure increases, the amount of feature information also increases, which actually increases the computational complexity and makes the model more sensitive. To overcome this kind of problem, a sampling process has been introduced:

- **Sub-sampling**

Sub-sampling or pooling is the procedure known as down-sampling the features to reduce dimensionality. Eventually it reduces the overall dimensionality and complexity. Four types of pooling operation are available:

- Max-Pooling
- Average Pooling.
- Mixed max-average pooling
- Gated max-average pooling.

Figure 3.4 illustrates a generalised pooling mechanism for a CNN model.

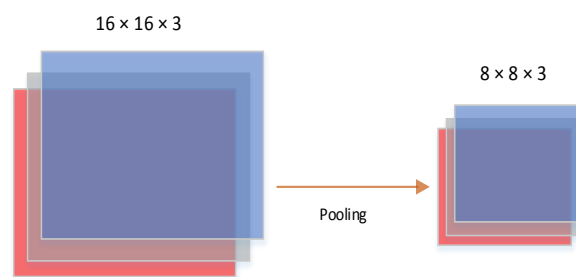


Figure 3.4: Pooling operation performed by 2×2 kernel

A DNN deals with a large number of neurons, which enables the network to take a direction where the network takes into consideration a large number of predictions.

This kind of situation provides very good performance in the training dataset and worse performance for the test dataset. This kind of problem is known as an over-fitting problem. To overcome this kind of problem the Drop-out procedure has been introduced. It is described in more detail below:

- **Drop-out**

Some of the neurons are randomly removed to overcome the over-fitting problem. In this procedure a few of the neurons are randomly dropped out (with some predefined probability) so that the network can learn more robust features. Figure 3.5 shows a simplified example of a drop-out mechanism. The right-hand side image shows that the network contains four hidden neurons 1 to 4; in the left-side image neurons 2 and 4 have been removed so that these two neurons do not have any effect on the network decision.

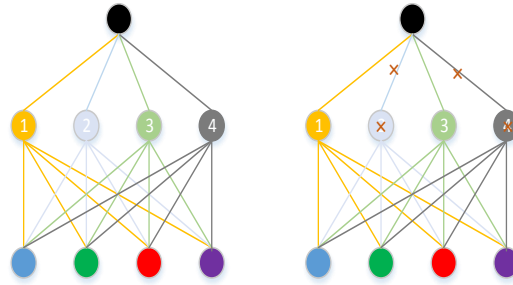


Figure 3.5: Drop-out

At the end of the network, all the neurons are arranged in a flattened way. The neurons of the flat layer are fully connected to the next layer and behave like a conventional neural network. Normally more than one fully connected layer is introduced. Consider the last layer as the "end" layer, then at the layer before the "end" layer; there must be at least one flat layer or fully connected layer. Then the end layer function can be represented as

$$\mathcal{F}_k^{end} = \sum_{j=1}^{end-1} w_{k,j}^{end} \mathcal{F}_g^{end-1} + \mathcal{B}_k^{end-1} \quad (3.10)$$

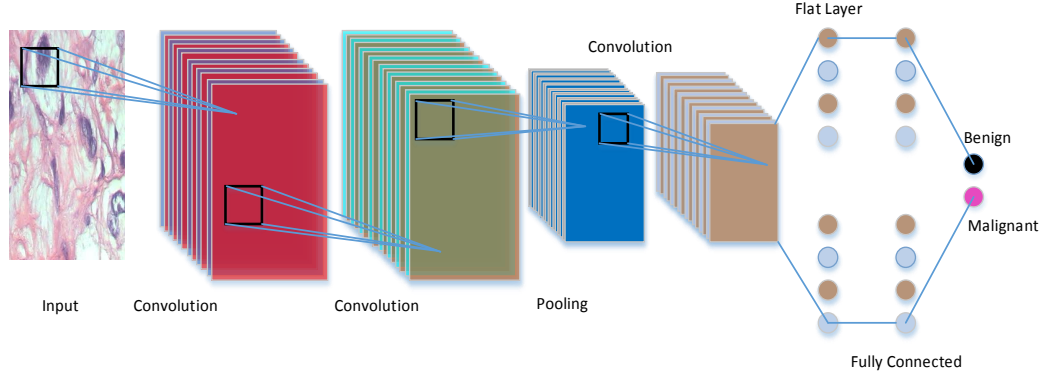


Figure 3.6: Workflow of a Convolutional Neural Network

Figure 7.7 depicts a generalised CNN model for image classification. The end layer can be considered as the decision layer.

Decision layer

In the decision layer Softmax-Regression techniques as well as the Support Vector technique are utilised.

- In the Softmax layer, the cross-entropy losses are calculated such as

$$L_k = -\ln(\bar{y}_k) \quad (3.11)$$

where \bar{y}_k can be written as:

$$\bar{y}_k = \frac{\exp(\mathcal{F}_k^{end})}{\sum_{k=1}^2 \exp(\mathcal{F}_k^{end})} \quad (3.12)$$

Here $k = \{1, 2\}$ where 1 is for Benign and 2 is for Malignant case. The value of L_k provide the final decision such as if $L_1 > L_2$ the network will produced Malignant output.

- **Support Vector Machine**

Instead of a Softmax layer, an SVM [208] layer can be used including the following

conditions. For a generalised case, let $x = x_1, x_2, \dots, x_n$ be the training data and $y = y_1, y_2, \dots, y_n$ be the corresponding label. If we consider that the data is linearly separable then the optimisation constraint is considered as $yW^T x \geq 0$. However, sometimes data is not linearly separable; in that case soft thresholding has been introduced and the constraint redefined as $yW^T x \geq 1 - \xi_i$. where $\xi_i = 0$. Now the optimisation problem is redefined as

$$\begin{cases} \min_{w, \xi_i} \frac{1}{2} W^T W + C \sum \xi_i \\ s.t. \xi_i \geq 1 - y_i W^T W x_i, \xi_i \geq 0 \forall_i \end{cases} \quad (3.13)$$

3.4.2 LSTM

While a CNN learns from scratch, an error signal is fed back to the input. In a Recurrent Neural Network, instead of learning from scratch the network learns from the reference point. The output of a particular layer is feed back to the input which works as the reference input. A generalised RNN model is presented in Figure 3.7. Let the sequence of input vectors be $\mathbf{X} = \{x_1, x_2, \dots, x_R\}$, the hidden state $\mathbf{H} = \{h_1, h_2, \dots, h_H\}$ and the output state $\mathbf{Y} = \{y_1, y_2, \dots, y_o\}$ where

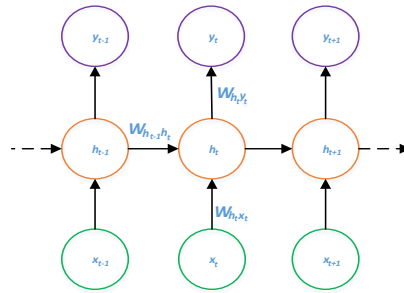


Figure 3.7: A generalised RNN model, where the RNN output is computed and the reference information passes through the hidden unit

$$y_t = \sigma(W_{h_t y_t} h_t + b_t) \quad (3.14)$$

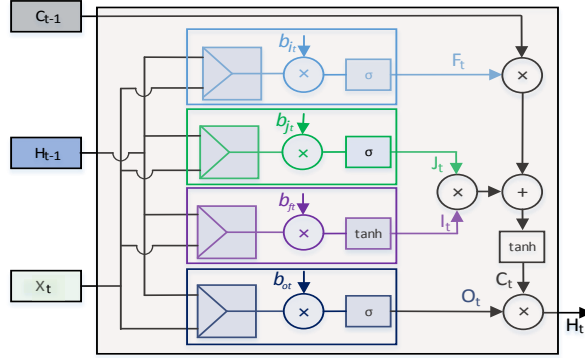


Figure 3.8: A generalised cell structure of an LSTM

Here, $W_{h_t y}$ represents the weight vector from the hidden unit h_t to the output unit y_t for the sequence t ; where h_t is defined as

$$h_t = \sigma(W_{h_{t-1} h_t} h_{t-1} + W_{x_t h_t} x_t + b_{h_t}) \quad (3.15)$$

Here, h_{t-1} represents the output of the hidden unit for the sequence $t-1$; $W_{h_{t-1} h_t}$ represents the weight vector from the hidden unit h_{t-1} to the hidden unit h_t for the sequence t ; b_{h_t} represents the bias; $W_{x_t h_t}$ represents the weight vector from the input sequence i_t to the hidden unit h_t .

A normal RNN suffers due to a vanishing-gradient probability. To overcome this problem, the Long-Short-Term-Memory (LSTM) architecture has been introduced by Hochreiter et al. [209]. One notable feature of the LSTM method is that it contains the "forget gate" through which the network controls the flow of information. Figure 3.8 represents the cell structure of an LSTM network. The main parameters of the LSTM network can be represented as:

$$i_t = \tanh(W_{x_t i_t} x_t + W_{h_{t-1} i_t} h_{t-1} + b_{i_t}) \quad (3.16)$$

$$j_t = \sigma(W_{x_t j_t} x_t + W_{h_{t-1} j_t} h_{t-1} + b_{j_t}) \quad (3.17)$$

$$f_t = \sigma(W_{x_t f_t} x_t + W_{h_{t-1} f_t} h_{t-1} + b_{f_t}) \quad (3.18)$$

$$o_t = \sigma(W_{x_t o_t} x_t + W_{h_{t-1} o_t} h_{t-1} + b_{o_t}) \quad (3.19)$$

$$c_t = c_{t-1} \odot f_t + i_t \odot j_t \quad (3.20)$$

$$h_t = \tanh(c_t) \odot o_t \quad (3.21)$$

f_t is the forget gate, i_t is the input gate, h_t provides the output information and c_t represents the cell state [210]. Here the weight matrix and bias vectors are $\mathbf{W}_{\times \times}$ and \mathbf{b}_{\times} .

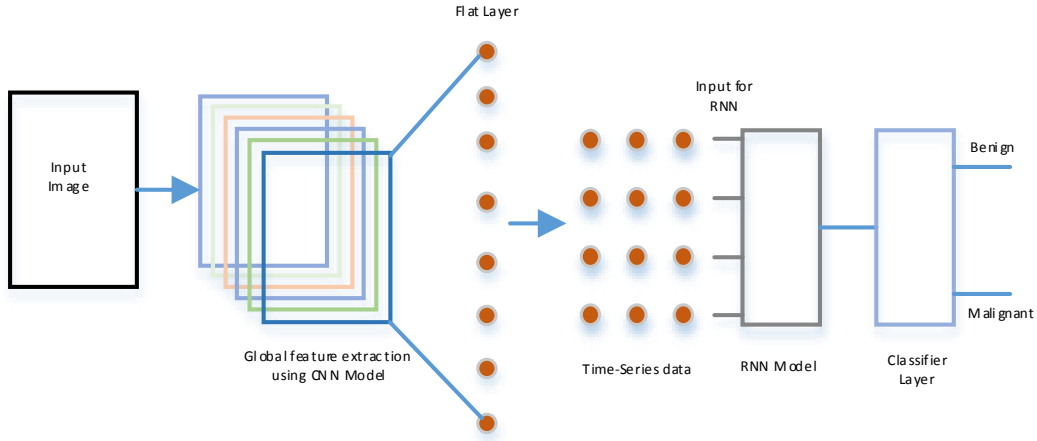


Figure 3.9: CNN and LSTM models combined

3.4.3 CNN-LSTM

A CNN has the benefit of extracting global information. On the other hand, an LSTM has the ability to take advantage of long-term dependencies of the data sequences. To utilise both these advantages, the CNN and LSTM models have been hybridised together for the classification [211], [212], [213].

From the output of the CNN model, it is difficult to generate an undirected graph to make the data into the time-series format, so that the network can extract the dependencies of the data. To do this we have converted the convolutional output (which is

2-dimensional) into 1D data. Figure 3.9 represents the basic structure of the LSTM and CNN model.

3.5 Proposed Models

We have utilised three different models for our data analysis. Model-1 utilises CNN techniques, Model-2 utilises the LSTM structure, whereas Model-3 employees both the CNN and LSTM structures together for the data analysis.

3.5.1 Model-1

In this method, the input image is convolved by a 3×3 kernel, and the output of each kernel is passed through an ReLU activation filter in layer C-1. Each kernel strides one step each time, and to keep the border information intact, we have added two extra rows and columns with a value of "0". This ensures that the newly created feature maps are also 32×32 in size. After the C-1 layer another convolutional layer named C-2 has been introduced, with the same kernel size 3×3 and an ReLU rectifier.

After the C-2 layer the pooling operation P-1 is performed with the kernel size 2×2 . As we have utilised a 2×2 kernel size, each of the feature maps decreases in size from 32×32 to 16×16 . After the P-1 layer another convolutional layer called C-3 has been utilised, with an ReLU rectifier. Each of the feature maps of the C-3 layer was 16×16 ; due to utilising the P-2 (Pooling layer of 2×2 kernel) layer the feature map is now 8×8 . After the C-4 layer another pooling operation has been performed named P-3 followed by a convolutional layer C-5. The output of the convolutional layer has been flattened. The C-5 layer contains 16 feature maps and each of the feature maps is 4×4 in size, so the Flattened layer contains 256 features. Twenty-five percent of the information has been dropped out in the dropout layer before sending them through the decision layer

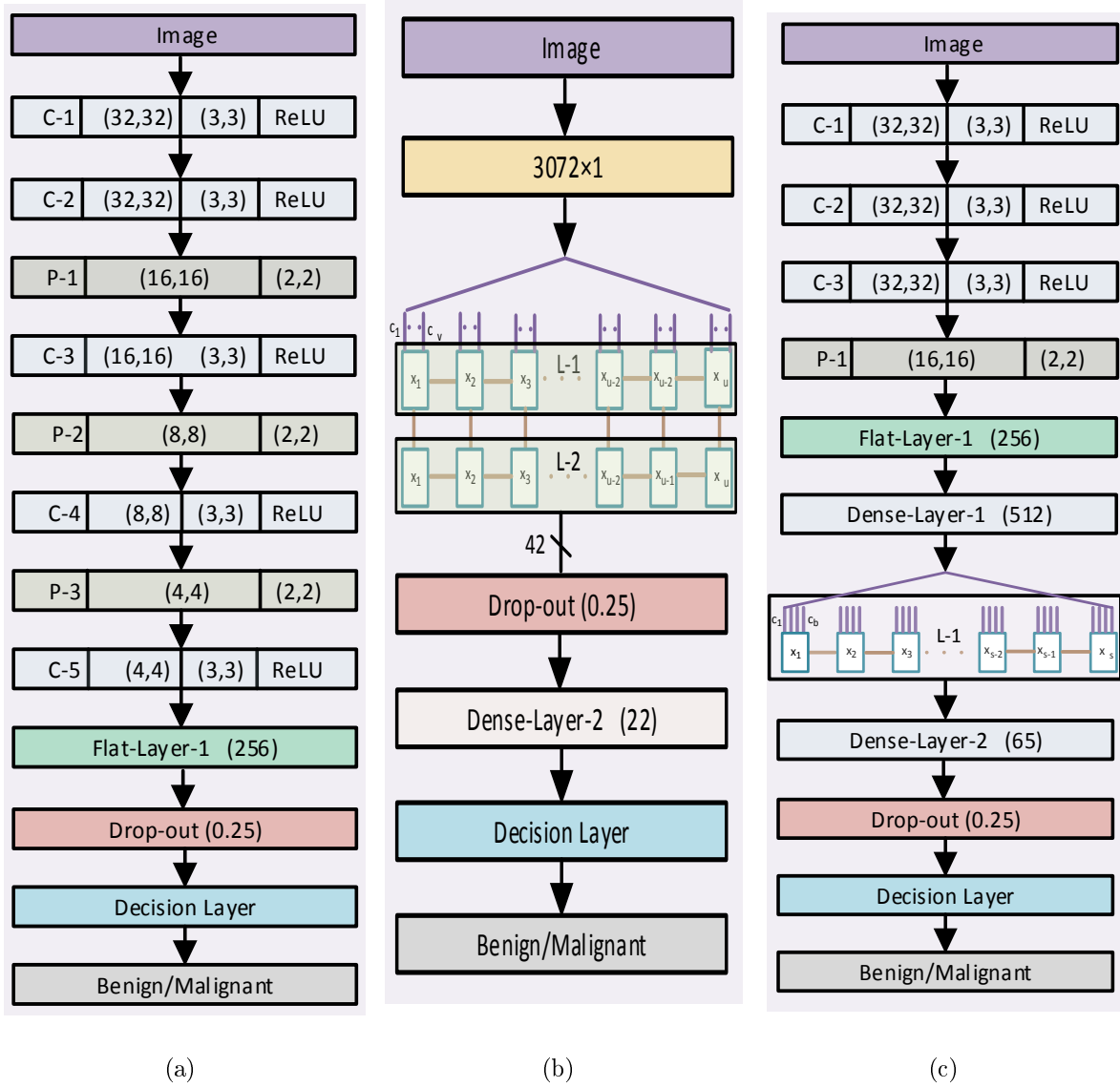


Figure 3.10: Conventional CNN, LSTM-based architecture (a,b), CNN-LSTM-based architecture (c)

(SVM/Softmax) to provide the benign or malignant decision.

3.5.2 Model-2

In the second model we utilised the LSTM method, which is a branch of the RNN model. Our input image is in two-dimensional format. To make it a suitable format for the LSTM model we have converted the data to 1-D data format, and the newly created data vector is 3072×1 in size, as our input data is $32 \times 32 \times 3$. This one-dimensional data has been converted to Time-Series data. To fit the 3072×1 into Time-Series data, we have created Time Steps (TS) data x_1 to x_u and the Input Dimension (ID) of each of the TS is a v such as c_1 to c_v , where $v \times u = 3072$. We stacked two LSTM layers consecutively, specifically L-1 and L-2. The output of the LSTM layer L-2 produces 42 neurons. The output of the LSTM layer is passed through the drop-out layer with a 25% probability. After the drop-out layer a dense layer has been introduced which contains 22 neurons. Finally a decision layer has been utilised to make the decisions about Benign and Malignant classes.

3.5.3 Model-3

In this model we have utilised both the CNN model and the LSTM model together. At first the input image is convolved by the convolutional layer C-1 with a 3×3 kernel along with a ReLU rectifier. This layer produces feature vectors and the size of each feature vectors is 32×32 . Consecutively there are another two layers, C-2 and C-3 placed one after another. After the layer C-3 one pooling layer named P-1 has been introduced with the kernel size 2×2 . As the pooling layer uses a 2×2 kernel, the output of P-1 produces a 16×16 kernel. After the P-1 layer a flat layer has been introduced, followed by a dense layer which produces 512 neurons. The output of this layer has been used as the input layer for the LSTM. As this layer contains a one-dimensional vector, we have converted this data into a time series. We have created TS data x_1 to x_s and each of the TS data

has contained an ID of size q such as c_1 to c_b where $s \times b = 512$. After the LSTM layer one dense layer of 65 neurons has been placed followed by a drop-out of 25% of the data. After that a decision layer has been placed which distinguishes the benign and malignant data.

3.6 Results and Discussion

We have utilised the BreakHis breast-image dataset for our experiment [205]. All the images of this dataset have been collected from 82 patients and the sample collection has been performed in the P&D Laboratory, Brazil. This dataset contains four group of images depending on the magnification factor $40\times$, $100\times$, $200\times$ and $400\times$. Each of the images of this dataset are RGB in nature and 760×460 pixel in size and they are elements of a particular set {Benign, Malignant}. Figure 3.11 shows the group-wise statistics as well as the overall statistics of this dataset.

As the Figure 3.11 shows, there are 7909 images where 2480 are Benign and the rest are Malignant, which indicates that almost 70.00% of the data are Malignant. For an individual magnification case, that is if we consider $40\times$, $100\times$, $200\times$ and $400\times$ individually, in all the cases almost 70.00% of the data are Malignant. This shows that this dataset is imbalanced, more specifically this dataset is more biased towards Malignant in terms of frequency.

3.6.1 Performance of different Models

Following a subsection analysing the performance of the algorithms based on parameters such as True Positive (TP/Sensitivity), False Positive (FP), True Negative (TN/Specificity), False Negative (FN), Accuracy, Precision, recall and Matthews Correlation Coefficient (M.C.C.). For the sake of comparison we have also performed all the experiments

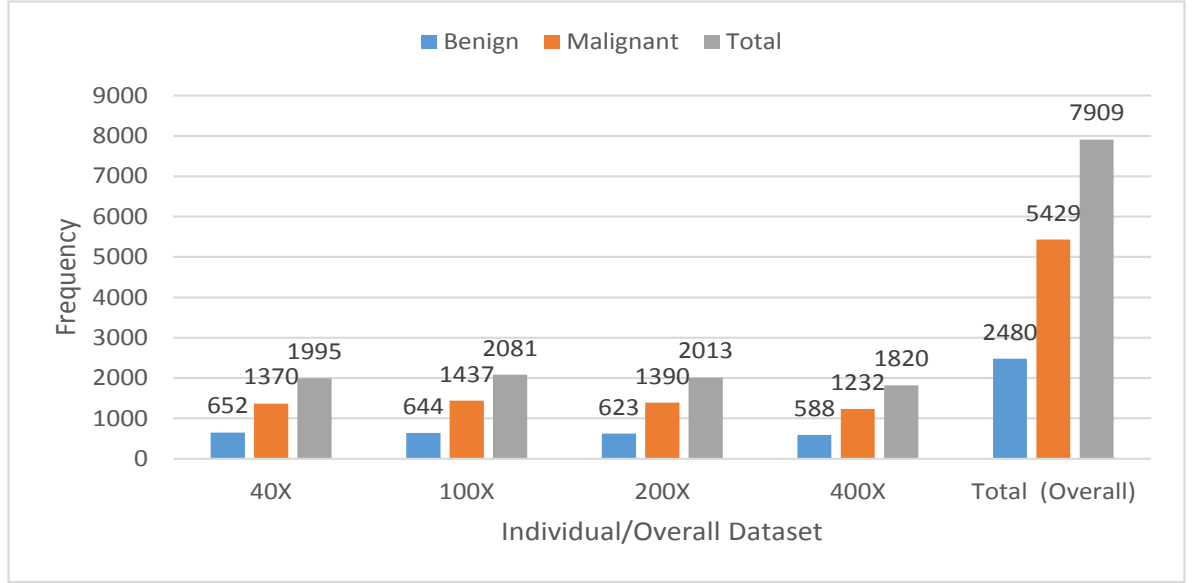


Figure 3.11: Statistical breakdown of the BreakHis dataset.

on the original images and this particular case is represented as (OI). When we utilised the KM algorithm we have fixed the cluster size (K) to 8, and when we utilized MS algorithm we have fixed the Bandwidth (BW) at 0.2

TP/FP/TN/FN performance

This subsection describes the True Positive (TP/Sensitivity), False Positive (FP), True Negative (TN/Specificity), False Negative (FN) performance from this experiment, and the data related to this experiment are presented in Table 3.2.

Table 3.2: Comparison of TN, FP, FN and TP values % for the different algorithms and different datasets

Dataset(\times)	Cluster	Decision Algorithm	Model-1				Model-2				Model-3			
			TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP
40 \times	MS	SVM	68.39	31.60	8.00	92.00	53.55	46.44	5.20	94.76	59.10	40.00	5.00	95.00
		Softmax	93.00	7.00	19.00	81.00	75.28	24.71	12.23	87.76	68.39	31.60	9.40	90.50
	KM	SVM	67.00	32.00	6.00	93.10	53.00	46.99	7.00	92.00	70.68	29.30	10.00	90.00
		Softmax	84.82	15.51	7.00	92.94	72.98	27.01	10.58	89.51	70.68	29.31	6.90	93.10
	OI	SVM	62.00	37.00	5.00	94.00	66.00	34.00	10.00	90.00	72.00	28.00	11.00	89.00
		Softmax	77.00	23.00	6.00	93.00	74.00	26.00	10.00	90.00	78.00	21.00	5.00	94.00
100 \times	MS	SVM	68.00	32.00	11.00	88.00	66.00	33.00	9.00	90.11	53.00	46.00	7.00	92.00
		Softmax	80.00	20.00	6.00	94.00	72.57	27.42	17.00	82.00	54.85	45.14	10.00	90.00
	KM	SVM	80.00	2.00	6.00	94.00	56.00	44.00	6.10	93.80	66.00	34.00	12.00	88.00
		Softmax	75.00	25.00	4.00	95.96	64.71	35.42	12.00	87.20	61.14	38.87	10.10	89.00
	OI	SVM	70.00	30.00	8.00	92.00	56.00	44.00	4.00	96.00	71.00	28.00	6.00	93.00
		Softmax	64.00	36.00	8.00	92.00	71.00	29.00	12.00	88.00	73.00	26.00	8.00	91.00
200 \times	MS	SVM	70.70	29.00	4.10	95.82	61.61	38.38	4.00	95.30	72.00	27.00	6.80	93.00
		Softmax	81.00	19.00	5.00	95.00	75.75	24.24	9.80	90.17	65.00	35.00	2.00	97.00
	KM	SVM	69.69	30.30	3.60	96.00	64.14	35.85	5.00	94.00	75.00	24.00	8.00	91.00
		Softmax	85.85	14.16	8.00	91.00	78.00	21.00	12.00	87.00	71.71	22.20	6.00	93.00
	OI	SVM	73.00	27.00	4.00	96.00	63.63	36.66	6.00	93.00	76.00	23.00	6.00	94.00
		Softmax	78.00	22.00	5.00	94.10	76.00	24.00	12.00	88.00	70.00	30.00	6.00	93.00
400 \times	MS	SVM	68.30	31.69	4.00	95.31	53.55	46.44	5.20	94.76	61.20	38.79	5.70	94.21
		Softmax	84.00	15.00	6.00	93.00	65.01	34.97	9.30	90.06	61.20	38.79	6.00	93.38
	KM	SVM	68.00	31.00	4.00	95.00	53.00	46.99	7.00	92.00	59.01	40.98	6.61	93.38
		Softmax	78.00	22.00	4.00	96.00	63.93	36.06	11.57	88.42	65.00	35.00	5.00	95.00
	OI	SVM	75.00	25.00	6.00	94.00	61.00	39.00	12.00	88.00	79.95	24.06	9.00	90.00
		Softmax	76.00	24.00	10.00	90.94	70.00	30.00	10.00	90.00	82.51	17.48	12.67	87.32

For the $40\times$ dataset the best True Positive (TP) value (95.00%) is achieved when Model-3 is utilised along with the MS cluster algorithm and the SVM classifier together. Model-2 also provides the same kind of TP value, 94.76%, when the MS and SVM algorithms are utilised together. In this particular case the TN values for Model-3 and Model-2 are 59.10% and 53.55%, respectively. However, when Model-1 is utilised in this particular scenario the TN value is 68.39% and the FP value is 31.60%. For the $40\times$ dataset, the best TN value is achieved when the MS cluster method and Softmax decision algorithm are utilised, and in this particular case the TP value is 81.00% for Model-1. When the original image (OI) is utilised, of the three models Model-1 provides the best TN and TP values, 78.00% and 94.00%, respectively. In this particular case a Softmax decision layer has been employed.

For the $100\times$ dataset the best TP value achieved 95.96% when we use KM clustering techniques and the Softmax decision algorithm together. In this particular case the TN value is 75.00% and the FP value is 25.00%. The best TN value, 80.20%, is achieved when we utilised the MS clustering algorithm and the Softmax algorithm together, and in this particular case the FP value is 19.80%. When the original image (OI) is utilised, the best TP value 93.00% is achieved when Model-3 along with with the SVM decision algorithm has been applied.

When we use the $200\times$ dataset the best TP value, i.e. 97.00%, is achieved when the MS clustering algorithm and the Softmax layer are utilised. However in this case the TN value is 65.00% and the FP value is 35.00%. When we use Model-1, MS, and SVM classifier together for the $200\times$ dataset, the TP value is 95.80% and in this case the TN and FP values are 70.70% and 29.00%, respectively. For the $200\times$ dataset the best TN value, 81.00 %, is achieved when the MS and Softmax algorithms are utilised with Model-1, and in that particular case the FP value is 19.00% , the TP is 96.00% and the FN value is 3.60%, respectively. A 96.00% TP value is achieved when the original image is utilised

along with Model-1. In this particular case the SVM Decision algorithm has been used.

For the $400\times$ dataset the best TP value achieved is 96.00% when KM and the Softmax layer along with Model-1 are utilised together. The best TP value achieved is 95.31% when we utilised Model-1 and the MS and SVM algorithms together. In this particular case the TN and TP values are 68.30% and 31.69% respectively. When we utilised the $400\times$ dataset the best TN value is 84.00%, achieved when we use the MS and Softmax algorithms (for Model-1) and the subsequent TP value is 93.00%. A 94.40% TP value is achieved when the original image is utilised along with Model-1 and the SVM Decision Algorithm. The best TP value is achieved when the original image is utilised along with Model-3 and the Softmax Decision Algorithm.

Accuracy performance

Figure 3.12 illustrates the Accuracy information for different models and different datasets. For the $40\times$ dataset the best Accuracy achieved is 90.00% when Model-1, the MS clustering method and a Softmax layer are utilised together. For the $40\times$ dataset and SVM classifier together, irrespective of the MS and KM clustering method, the Accuracy performance is almost the same at 86.00%. For the $40\times$ dataset, of all three models, Model-1 gives the best performance for all cases irrespective of the cluster method as well as the classifier method. When we use the $40\times$ dataset the best Accuracy performance is achieved when Model-1 and a Softmax layer are combined.

For the $100\times$ dataset and the MS cluster method along with the SVM method, Model-2 provides the best performance, 83.13%, and this same kind of Accuracy performance, 83.00%, is shown by Model-1. When the KM cluster and SVM classifier are used together, Model-1 provides 84.87% Accuracy followed by Model-2 (82.97 %) and Model-3 (81.78%). When the Softmax classifier is utilised Model-1 elicits the best Accuracy performance irrespective of the clustering method, whether the MS or KM cluster method is employed.

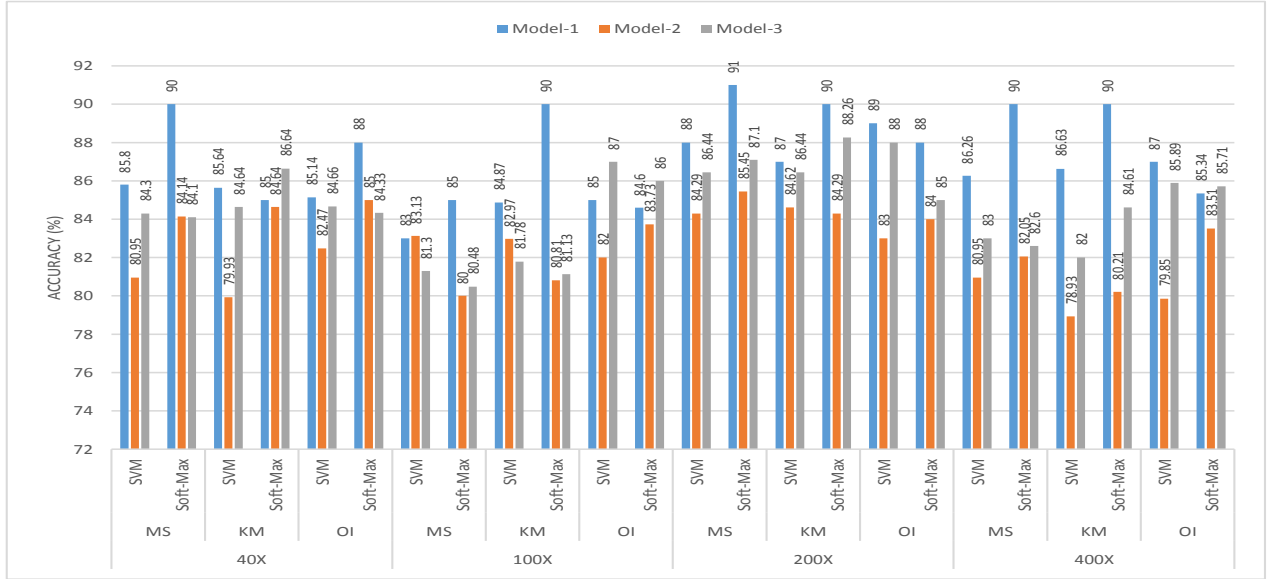


Figure 3.12: Comparison of Accuracy between Model-1, Model-2 and Model-3

Model-1 and Model-3 provide the same kind of Accuracy performance of around 81.00% when we use the Softmax classifier, and this result remains the same whether we use the MS or KM cluster algorithm. When we use original images, of the three models, Model-3 provides the best Accuracy performance, 87.00%, where SVM classifier layers have been utilised.

When we use the 200 \times dataset and the MS clustering algorithm, for all the models the Softmax classifier performs better than the SVM classifier. The best Accuracy of 91.00% is achieved when we use Model-1. For the K-M cluster algorithm, the Softmax classifier provides better performance than the SVM classifier. When we use the original images the best Accuracy is achieved when Model-1 has been utilised along with an SVM classifier layer.

For the 400 \times dataset with the Softmax classifier, the best Accuracy performance (90.00%) is achieved when we utilised Model-1 irrespective of the MS or KM algorithm. When we utilised the SVM algorithm Model-1, provides better Accuracy (around 82.26%),

than Model-2 and Model-3. For the $400\times$ dataset Model-1 shows the best performance when we utilized an SVM layer.

Precision performance

Figure 3.13 shows the Precision information for different models and different datasets. For the $40\times$ dataset the best Precision performance (96.00%) is achieved when the MS cluster algorithm and a Softmax layer are utilised with Model-1. When the KM clustering algorithm and Softmax classifier are utilised together, the best Precision (94.00%) is achieved when we employed Model-1. Interestingly, when the KM clustering method and Softmax layer are utilised both Model-2 and Model-3 give a similar Precision of 89.00%. The worst Precision value (80.00%) is achieved when we utilise the KM clustering algorithm and SVM classifier with Model-2. Overall, for the $40\times$ dataset, the SVM classifier provides the worst performance when the Softmax layer is utilised. When we utilise original images the best Precision value (92.00%) is achieved for Model-3 along with a Softmax decision layer.

For the $100\times$ dataset the best Precision (91.00%) is achieved when we used the KM clustering algorithm along with the Softmax layer with Model-1. In this particular situation, Model-2 and Model-3 provide 86.00% and 85.00% Precision, respectively. For KM clustering and SVM classifier both Model-1 and Model-2 achieve 87.00% Precision. When the MS clustering method is implemented the best performance is achieved when Model-3 is used along with the Softmax layer. For the MS clustering method, Model-1 and Model-3 provide similar levels of Precision. When we utilise original images the best Precision value (89.00%) is achieved for Model-1 along with a Softmax decision layer.

For the $200\times$ dataset the best Precision (93%) is achieved when the KM clustering method and a Softmax layer and Model-1 algorithm are utilised together. In this

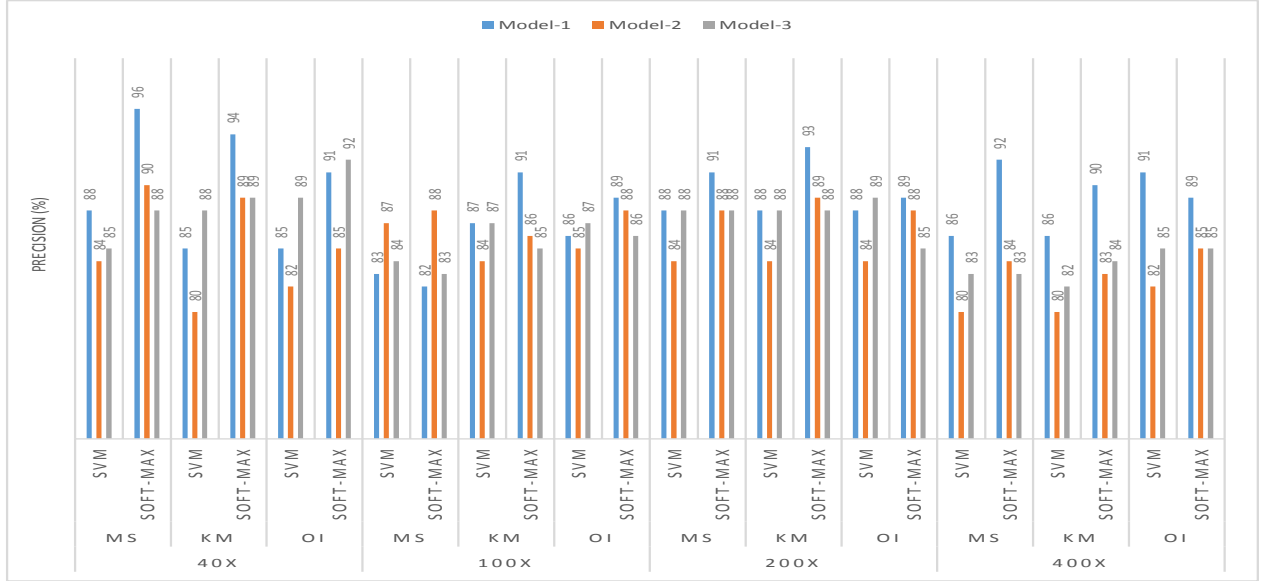


Figure 3.13: Comparison of Precision between Model-1, Model-2 and Model-3

particular case Model-2 and Model-3 provide a similar Precision of 89.00% and 88.00%, respectively. For the MS cluttering algorithm, the best Precision, 91.00%, is achieved when Model-1 is utilised. For the KM clustering algorithm and the SVM method the Precision achieved is 88.00%. When we utilise original images the best Precision value (89.00%) is achieved for Model-1, and this result is true for both the SVM as well as the Softmax decision layer.

For the $400\times$ dataset, the best performance is achieved when the MS clustering method along with the Softmax layer is utilised; Model-1 provides the best Precision (92.00%). In this particular case Model-2 and Model-3 provide 84.00% and 83.00% Precision, respectively. With the KM clustering and the Softmax layer together the Precision value is 90.00%. Overall, the Softmax layer provides the best Precision values. A 91.00% Precision value is achieved for Model-1 and the SVM Decision-layer algorithm when an original image has been provided as input.

F-Measure performance

Figure 3.14 shows the F-Measure information for different models and different datasets. For the $40\times$ dataset when the KM clustering method with the Softmax layer is used, an F-Measure 93.00% value is achieved when Model-1 is utilised. In that particular scenario Model-2 gives a 91.00% F-Measure and Model-3 an 89.00% F-Measure. For the MS clustering algorithm and SVM classifier algorithm, Model-1 and Model-2 provide 90.00% F-Measure values. In this particular clustering algorithm, when the Softmax layer is employed all the models provide the same performance, around 89.00%. A 93.00% F-Measure value is achieved when we utilise Model-3 along with the Softmax algorithm and original image.

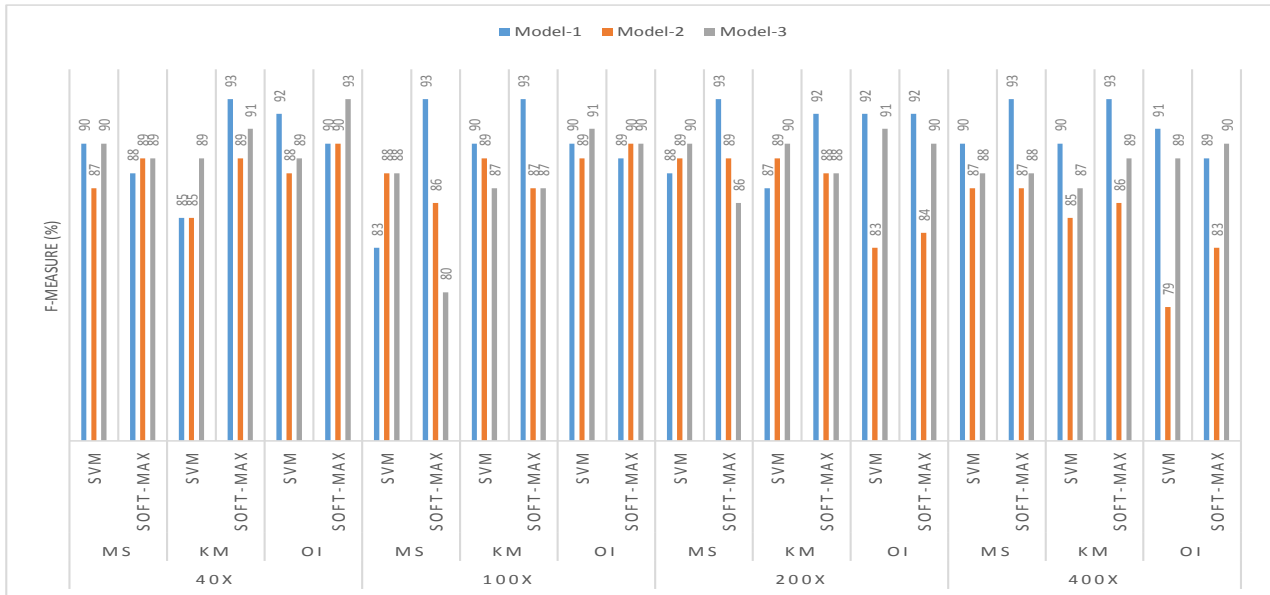


Figure 3.14: Comparison of F-Measure between Model-1, Model-2 and Model-3

For the $100\times$ dataset Model-1 provides the best F-Measure of around 93.00% when the Softmax layer algorithm is employed; this performance is true for both the MS and KM clustering methods. When KM clustering and the Softmax layer are combined together Model-2 and Model-3 provide the same F-Measure of 87.00%. When the KM clustering

method is utilised with the SVM classifier, Model-1 gives a 90.00% F-Measure while Model-2 and Model-3 provide 89.00% and 87.00% F-Measure values, respectively. When the MS clustering algorithm is combined together with Model-2 and Model-3 provide the same F-Measure of 88.00%, and with this particular scenario Model-1 provides an 83.00% F-Measure. A 91.00% F-Measure value is achieved when we utilise Model-1 along with the SVM algorithm at the decision layer and provide original image. In this particular case when we utilise the Softmax layer both Model-2 and Model-3 provide similar F-Measure values.

For the $200 \times$ dataset, the best F-Measure of 93.00% is provided by Model-1 when the MS algorithm and Softmax layer are combined. However, when the KM cluster is utilised along with the Softmax layer the F-Measure is 92.00%. In this particular scenario, both Model-2 and Model-3 provide a similar F-Measure value of 88.00%. When KM clustering and the SVM algorithm are utilised together Model-3 provides an 90.00% F-Measure, Model-2 provides a 89.00% F-Measure, and in this particular case Model-1 provides an 87.00% F-Measure. When SVM and the Softmax layer are used together Model-1, Model-2 and Model-3 provide 88.00%, 89.00% and 90.00% F-Measure, respectively. A 92.00% F-Measure value is achieved when we utilise Model-1 and original image, and this is true for both the Softmax and SVM algorithms.

For the $400 \times$ dataset Model-1 provides the best F-Measure value of 93.00% irrespective of the clustering method. For the KM clustering algorithm and SVM algorithm, the F-Measure values are 90.00%, 85.00% and 87.00% for Model-1, Model-2 and Model-3 respectively. When the MS clustering method and SVM algorithm are utilised together Model-1, Model-2 and Model-3 provide 90.00%, 87.00% and 88.00% F-Measure values, respectively.

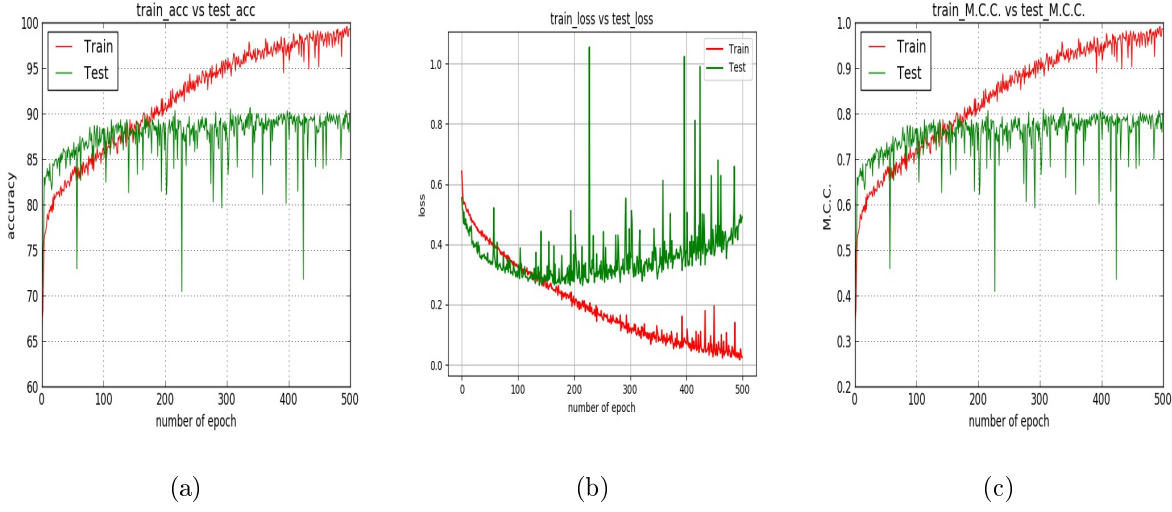


Figure 3.15: Accuracy, Loss and M.C.C. values for Model-1 when we utilised the $40\times$ dataset MS and Softmax together

Accuracy, loss and M.C.C performance at different epochs

The best Accuracy performance is achieved when we utilised Model-1 along with MS clustering and the Softmax layer on the $40\times$ dataset. Figure 3.15 a, b and c represent, respectively the Accuracy, Loss and M.C.C. values for this particular situation. Initially the Test Accuracy shows better performance than the Train Accuracy. Up to around epoch 180, the Train Accuracy is better than the Test Accuracy. After the epoch 180 the Train Accuracy exhibits superior performance than the Test Accuracy.

After epoch 300 the Train Accuracy remains constant at about 90.00%. Interestingly, after around epoch 180 the Train Accuracy outperforms the Test Accuracy, after around epoch 180 the difference in Accuracy performance between the Train and Test increased, with the Test remaining constant.

Model-2 provides the best Accuracy with the $200\times$ dataset and the MS algorithm and Softmax layer. Figure 3.16 shows the Accuracy, loss and M.C.C. values for this particular case for epoch 500. On virtually every occasion the Train Accuracy performance is better than that of the Test Accuracy. After about epoch 100 the Test Accuracy almost

remained constant, however the Train Accuracy continuously increased, and after epoch 300 the Train Accuracy touches 100% and remains constant throughout the epochs. Figure 3.16 b shows that the Train loss continuously decreases and the Test Accuracy steadily increases. As the epoch progresses the gap between the train loss and test loss continuously increases. The test M.C.C. remains almost constant around 0.73 while the train M.C.C. value continuously increases and touches 1 and remains constant.

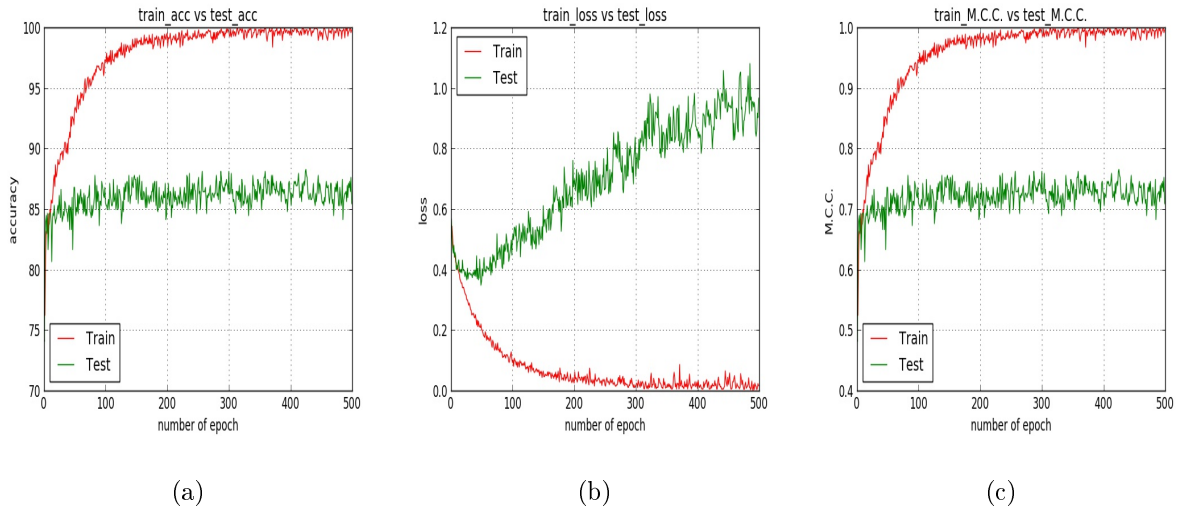


Figure 3.16: Accuracy, loss and M.C.C. values for Model-2 when we utilised the $200\times$ dataset, MS and Softmax together

Model-3 is the most accurate with the $200\times$ dataset and the KM and Softmax layer. Figure 3.17 shows the Accuracy, loss and M.C.C. values for this particular case for epoch 500. Figure 3.17 (a) shows that the Train Accuracy is almost always higher than the test Accuracy. The difference between the Train Accuracy and the Test Accuracy increases with the epoch up to around epoch 100. After epoch 100 the Test Accuracy remains constant at around 88.00% and the Train Accuracy remains constant at 100.00%. For the loss performance, the Test loss reduces as the epoch progresses on and the Train loss value remains virtually constant. The M.C.C. value for the test (around 78.00%) remained constant after around epoch 20. The train M.C.C. value touched the highest

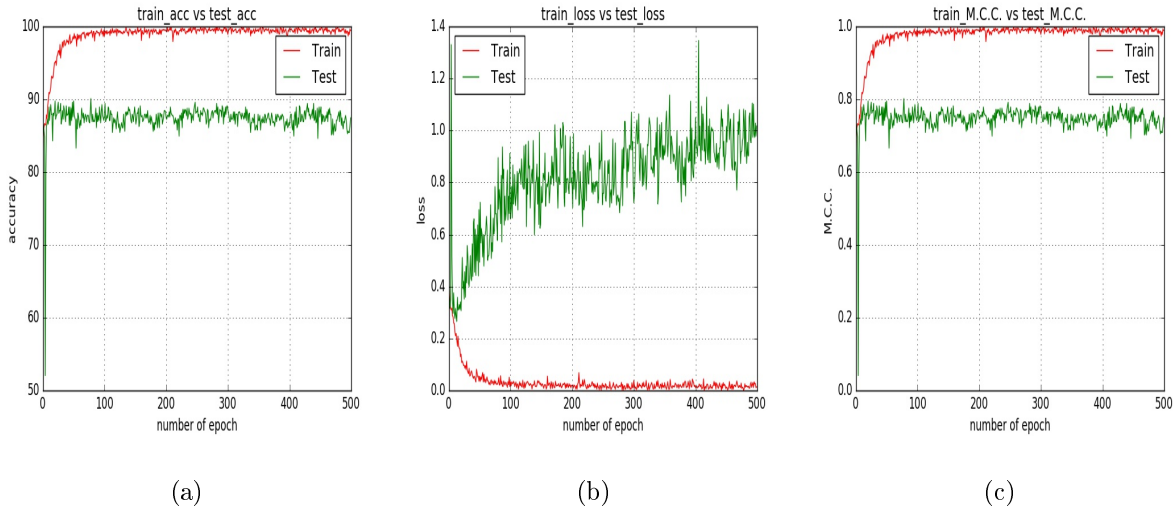


Figure 3.17: Accuracy, loss and M.C.C. values for Model-3 with the $200\times$ dataset, KM and Softmax together

value, of 1.00, after around epoch 100.

3.6.2 Effect of TS and ID

TS and ID have an effect on LSTM performance. In this subsection we analyse the effect of the TS and ID values with reference to Accuracy, average time and required parameters for Model-2.

Table 3.3: Average time and Parameters for various TS and ID

TS	ID	Average Time (s)	Parameters
24	128	191	58280
32	96	240	52904
48	64	346	47528
64	48	438	44840
96	32	636	42152
128	24	822	40808

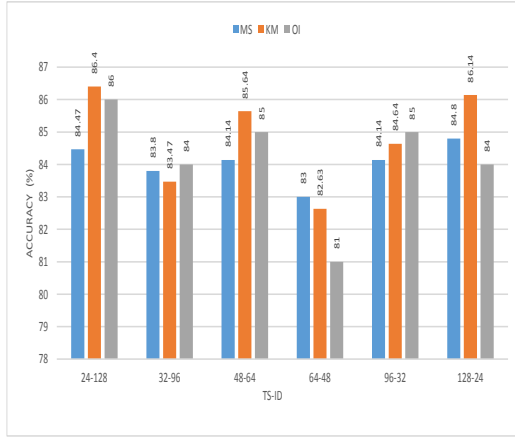
Table 3.3 summarises the average time and parameters required for Model-2 perfor-

mance with different combinations of TS and ID . When TS and ID are fixed at 24 and 128, respectively, the required average time is 191 seconds and a total of 5808 parameters are required. This table also exhibits a very interesting behaviour. As we increase the value of TS and reduce the value of ID, the number of required parameters to execute the CNN model has fallen. However, the time required to execute the model increased.

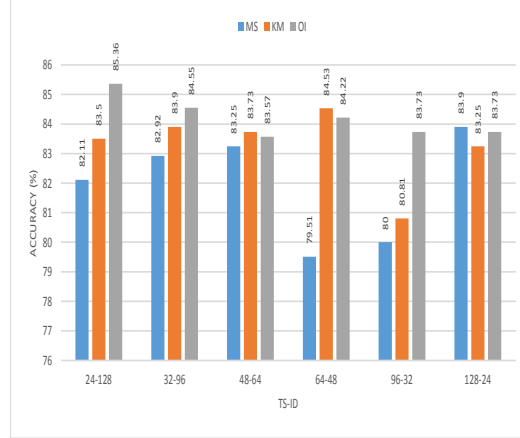
For the $40\times$ dataset, Figure 3.18 (a) shows the Accuracy where the TS and ID values have been varied. When TS and ID are fixed at 24 and 128 respectively the obtained Accuracy for the MS, KM and OI methods were 84.47%, 86.4% and 86.00% respectively. Figure 3.18 (b) displays the Accuracy performance on the $100\times$ dataset with different TS and ID values. Where TS=24 and ID=128, 85.36% Accuracy is achieved when the original image is utilised. When the TS value is fixed at 128 and ID are fixed at 24, the MS method provides Accuracy at 83.90% . For the $200\times$ dataset, 86.94% Accuracy has been achieved using the MS method with the TS and ID values 64 and 96, respectively. When TS and ID is fixed at 128 and 24, respectively, the Accuracy was 87.00%. For the $400\times$ dataset 84.24% Accuracy is achieved when the MS method is utilised, where TS is fixed at 64 and ID is fixed at 48.

3.6.3 The effect of Cluster size (K) and Bandwidth (BW)

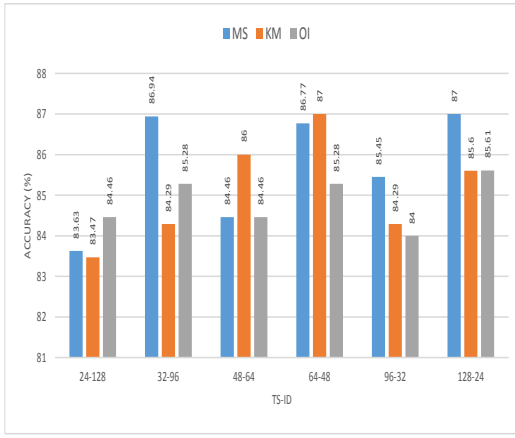
For the local partitioning we have utilised KM and MS algorithms. The cluster size of the KM method and the Bandwidth (Neighbour size) of the MS method largely control the performance of the clustering. In this subsection we investigate how these two parameters affects the overall performance which has been presented in Table 3.4. For this particular analysis we have only considered the $200\times$ dataset and Model-1. We have utilised the values of K equal to 8, 16 and 24. As the value of K increases, the TP value also increases. This indicates that with increasing K, the model performs in a specific way. Among the three values of K the best TN value (85.85%) is achieved when we utilise K=8. Overall



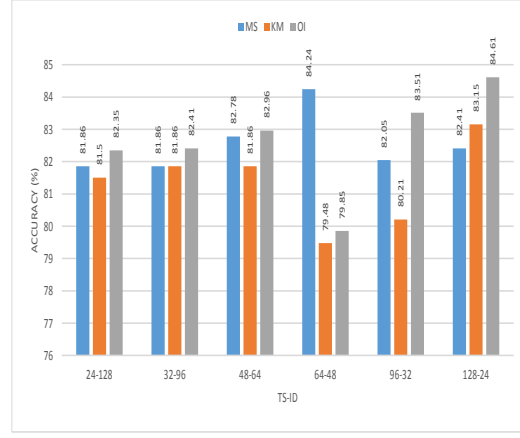
(a)



(b)



(c)



(d)

Figure 3.18: Figures a, b, c and d represent the Accuracy of the $40\times$, $100\times$, $200\times$, $400\times$ datasets for Model-2 with varying TS and ID

the best Accuracy is achieved when we utilised K=8 which is slightly better than with K=24.

Table 3.4: Effect of the cluster size (K) and the Bandwidth (BW)(%)

		TN	FP	FN	TP	Precision	F-Measure	Accuracy
KM	K=8	85.85	14.16	8.00	91.00	91.00	92.00	90.00
	K=16	77.00	23.00	06.00	94.00	90.00	92.00	88.90
	K=24	77.00	23.00	05.00	95.00	90.00	93.00	89.75
MS	BW=0.2	81.00	19.00	5.00	95.00	93.00	93.00	91.00
	BW=0.4	70.00	30.00	04.00	96.00	87.10	91.00	87.00
	BW=0.6	76.00	24.00	06.00	94.00	89.00	91.00	87.00

For the MS method the obtained Precision values are 93.00%, 87.10% and 89.00%, respectively, for BW equal to 0.2, 0.4 and 0.6, also respectively. The best Accuracy performance (91.00%) is achieved when we utilised BW=0.2. For both BW equal to 0.4 and 0.6 the obtained Accuracy was 87.00% which is less than when BW is equal to 0.2.

3.7 Recent findings for Breast-Image Classification based on DNN

DNN methods have been implemented for breast-image classification with some success. Table 3.5 shows recent findings of breast-cancer image classification based on the DNN method used for Histopathological images (other than the BreakHis dataset). The best Accuracy performance of 92.45% is achieved by B. Bejnordi [214].

However, we cannot exactly compare our performance with this existing finding because of the different datasets. We have compared our findings with the findings based on the BreakHis dataset which are presented in Table 3.6. F. Spanhol classifies the BreakHis

Table 3.5: CNN and Histopathological findings

Authors	Data Set	Method	Augmen- tation	No of Class	Accuracy %	Sensitivity %	Recall %	ROC
T. Araujo et al. [204]	[215]	CNN	YES	2	80.60	70.00	—	—
T. Araujo et al. [204]	[215]	CNN+SVM	YES	2	83.20	80.00	—	—
B. Bejnordi	BREAST	CNN	YES	—	92.00	—	—	92.00
B. Bejnordi [214]	[214]	CNN	YES	—	92.45	—	—	—

dataset into benign and malignant classes using a CNN model and a few other models. Their CNN model is similar to the Alexnet CNN architecture and their finding (best one) has been listed in Table 3.6. In our experiment for the $40\times$ dataset, we obtained 90.00% Accuracy whereas Spanhol et al. [7] obtained 90.40%. However, for the $100\times$, $200\times$, and $400\times$ datasets the best achieved accuracies in our experiment are 90.00, 91.00 and 90.00%, respectively, which is better than the findings of Spanhol et al. [7]. Apart from this, Spanhol et al. [7] have no information about the sensitivity, Precision, recall and M.C.C. values. In this work we have explained those issues in detail. The original image of the BreakHis dataset is $760\times460\times3$ pixels, and when Spanhol et al. [7] use this image they convert it to $350\times230\times3$ pixels. However, we have utilised a $32\times32\times3$ pixel image which has reduced the computational latency [7]. K. Dimitropoulous et al. [216] utilised the Grassmannian Vector of Local Aggregated Descriptor (VLAD) method for the BreakHis dataset classification. Their finding is comparable to our finding. However, in their paper they did not utilise the DNN models. Also, they do not describe the sensitivity, specificity, F-Measure and M.C.C. values, whereas we have explained those terms explicitly.

Table 3.6: Comparing Accuracy (%) in different models

	40×	100×	200×	400×
CNN [7]	90.40	87.40	85.00	83.80
VLAD [216]	91.80	92.10	91.40	90.20
PFTAS [216]	83.80	82.10	85.10	82.30
ORB [216]	74.40	69.40	69.60	67.60
LPQ [216]	73.80	72.80	74.30	73.70
LBP [216]	75.60	73.20	72.90	73.10
GLCM [216]	74.70	78.60	83.40	81.70
CLBP [216]	77.40	76.40	70.20	81.80

3.8 Conclusion

The judgement about benign and malignant status from digital Histopathological images is subjective and might vary from specialist to specialist. CAD systems largely help to make an automated decision from the biomedical images and allow both the patient and doctors to have a second opinion. A conventional image classifier utilises handcrafted local features from the images for the image classification. However, the recent state-of-the-art DNN model mostly employs global information using the benefit of kernel-based working techniques, which act to extract global features from the images for the classification. Using this DNN model, this chapter has classified a set of Breast-Cancer images (BreakHis dataset) into benign and malignant classes.

Images normally preserve some statistical and structural information. In this chapter, to extract the hidden structural and statistical information, an unsupervised clustering operation has been done and the DNN models have been guided by this clustered information to classify the images into benign and malignant classes. At the classifier stage

both Softmax and SVM layers have been utilised and the detailed performance has been analysed. Experiments found that the proposed CNN-based model provides the best performance other than the LSTM model and the combination of LSTM and CNN models. We have found that, in most cases, Softmax layers do perform better than the SVM layer.

Most of the recent findings on the BreakHis dataset provide information about the Accuracy performance but do not provide information about the sensitivity, specificity, Recall, F-Measure and M.C.C.; however, we have explained these issues in detail. The best specificity, sensitivity, Recall and F-Measure are 96.00%, 93.00%, 96.00% and 93.00% respectively. Of these issues, this chapter has explained how the Accuracy, M.C.C. and loss values change with different epochs.

Providing a definite conclusion about the biomedical situation needs to be considered as it is directly related to the patient's life. In a practical scenario, the classification outcome of the BC images should be 100.00% accurate. Due to the complex nature of the data we have obtained 91% Accuracy, which is comparable with the most recent findings. There are a few avenues for obtaining more reliable solutions such as the following:

- Each Histopathological image contains cell nuclei, which provide valuable information about the malignancy. So the DNN model guided by the cell nuclei orientation and position can improve the performance, since it provides more objective information to the network.
- As our dataset is comparatively too small to be used with a DNN model, in future the following two cases can be considered:

1. Data Augmentation
2. Transfer Learning

with some fine local tuning.

-
- Locally hand-crafted features also provide valuable information. So parallel feeding of the local data along with the raw pixels could improve the model's performance with reference to Accuracy.

Chapter 4

Histopathological Breast-Image

Classification With Image

Enhancement by Convolutional Neural

Network

4.1 Abstract

Finding malignancy from Histopathological images is always a challenging task. So far research has been carried out to classify Histopathological images using various techniques and methods. Recently, the state-of-the art Convolutional Neural Network (CNN) has largely been utilised for natural image classification. In this chapter, using the advancement of CNN techniques, we have classified a set of Histopathological Breast images into

Published as: A. A. Nahid, F. B. Ali, and Y. Kong, “Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network”, in *2017 20th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 1-6, Dec. 2017.

Benign and Malignant classes, which can save doctors and physicians time and also allow patients a second opinion about the disease.

4.2 Introduction

Cancer is a combination of a few diseases. The cells of a body maintain a lifestyle where a few die and a few cells grow and maintain a check and balance. This is the normal phenomenon of the cells of the body, however different situations can happen. Sometimes cells grow with no constraint and this growing can persist. This can lead to some unwanted situations and create cancer in the body. Almost all cancer-affected people lead unbearable and miserable life conditions.

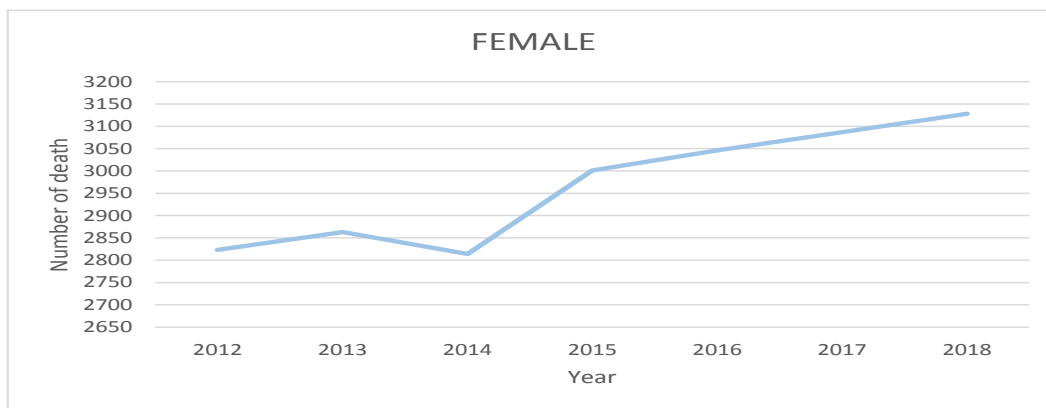


Figure 4.1: Number of Female deaths in Australia since 2012 due to breast cancer

The people of the whole world face a serious threat from cancer, as the human body is a combination of cells, and cancer starts from the cells. So cancer can be created in any part of the body and later can distribute to any other part of the body. Different kinds of cancer exist, such as Liver, Skin, Breast etc. Statistics show that, around 8.8 million died due to cancer in 2015 [217]. Of all the cancers, women are more vulnerable than men to breast cancer due to their physical anatomy. Figure 5.2 shows the Breast Cancer

statistics of Women for the last seven years in Australia (population 20-25 million). This graph shows that almost every year the number of female deaths due to breast cancer increased. Though this is an example from Australia this might be considered as a picture for the whole world.

The creation of a cancer depends on various issues, such as family history, DNA structure, lifestyle, smoking etc. When cancer is once created inside the body the only way of recovery from the cancer is proper treatment. The identification of the Cancer is a prerequisite for cancer treatment. Among other methods, investigation of biomedical images always helps doctors diagnose cancer. The biomedical photography techniques can be classified into Invasive (Biopsy Images) and non-invasive (such as Mammogram). Histopathological images are collected from the biopsy images. This Histopathological image investigation is always very challenging and time-consuming, and requires expert knowledge to reach a final decision. Sometimes experts fail to make a final decision.

Different research groups have investigated the analysis of Histopathological images using different mathematical models and techniques. However the most recent state-of-the art CNN techniques have been largely utilised for image classification techniques. Advanced engineering has been utilised with the CNN model for Biomedical image classification.

The CNN model is an advanced utilization of the Neural Network. The history of the Neural Network is a long one. The working root of the Neural Network relies upon the perceptron algorithm, which models the working principle of the human brain [45]. The Convolutional Neural Network is one branch of Neural Networks. Convolution Neural Network techniques re-gained a focus after the great work of Hilton in his paper [218]. The Convolutional Neural Network (CNN) is a very recent concept, and has made revolutionary changes in the data processing task, specially after the work of Alex Krizhevsky in his model AlexNet [219]. This CNN model has been used for the image classification

task. After the model AlexNet, the GoogleNet model was introduced [220], where the authors use deeper and wider networks for image classification. They claim that widening and deepening the size of the network can improve the performance of the classification task while keeping constraints on the network parameters.

Conventional Supervised image classifier techniques such as Support Vector Machine (SVM), Random Forest (RF) etc. or even the conventional Neural Network (NN), largely depend on extraction of locally handcrafted feature information which follows some predefined mathematical or logical tools. However, the Convolutional Neural Network depends on extraction of global as well as local features utilizing kernel methods. The extraction and modeling requires more time to classify the data using a CNN architecture, but it provides some excellent performance. To limit the complexity within the constraints two layers like Pooling and Drop-out can be used. These two layers allow the network to reduce the complexity.

Normally image classification requires a few preprocessing steps. The performance of the image classification depends on the nature of the images. Histopathological images suffer from color inconsistency due to the following issues:

- Chemicals.
- Stains.
- Lighting

Of the previous three issues, lighting variation causes an uneven distribution of the illumination. There are a few methods available for image illumination correction, among them the Retinex operation which performs a non-linear transform to correct the illumination. In this chapter we enhance the image utilizing the Retinex filter, and then classify the images into Benign and Malignant categories. We have organized the chapter as: Section 4.2 describes recent related work, Section 4.3 describes the overall model for image classification along with description of the Retinex filter, Section 4.4 describes the basic Convolutional Neural Network along with our three models for the image classifi-

cation task. Section 4.5 describes the results and analyzes the performance of the overall classification system. Section 4.6 concludes this chapter.

4.3 Overall Architecture for Classification

Conventional Supervised image classification is always a challenging task which follows some predefined steps such as the selection of the image database, preprocessing of the images, crafting the features, and selection of the classifier model. However the state-of-the-art CNN method mostly extracts the features globally and utilises both the local and the Global features for the image classification. Raw images always suffer from local statistics, noise and illumination variations due to the variation of the image sources and environment. Instead of directly using the raw images, we firstly normalize the image by applying the Multiscale Retinex algorithm to improve the local contrast and illumination variation. The overall image-classification architecture is presented in Figure 5.2.

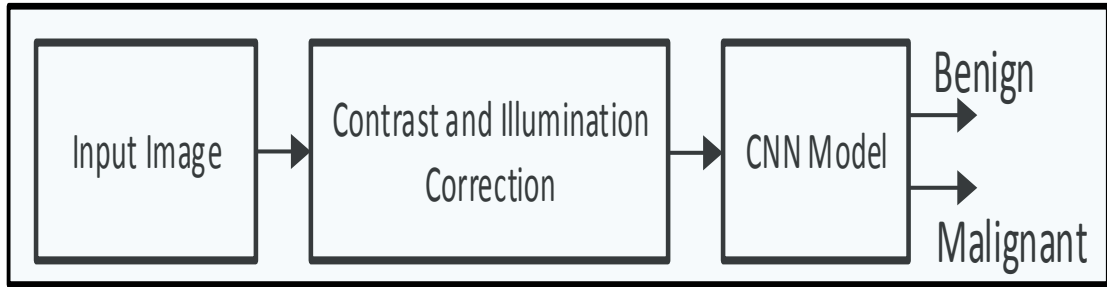


Figure 4.2: Overall image-classification model

The following subsection briefly describes the Retinex algorithm used for the image preprocessing steps.

4.3.1 Retinex Algorithm

The word Retinex is a combination of the words Retina and Cortex for the algorithm which was first proposed by E. Land in 1977 [221]. Good visual perception of a scene can be achieved based on the Retinex algorithm, where non-linear transforms are used to improve the color consistency. Single-scale Retinex can be expressed as

$$\mathcal{R}_i = \alpha \{ \log_{10} I_i(x_i, y_i) - \log_{10} [I_i(x_i, y_i) \times \mathcal{F}(x_i, y_i)] \} - \beta \quad (4.1)$$

\mathcal{R}_i is the Retinex image where I_i is the input image of channel i . \mathcal{F} is the normalized kernel which can be represented as

$$\mathcal{F}(x, y) = k \exp \left[-\frac{(x_i^2 + y_i^2)}{\sigma^2} \right] \quad (4.2)$$

where α and β are a scaling factor and an offset parameter respectively. For simplicity of calculation we will avoid the α and β terms. So equation (1) can be written as

$$\mathcal{R}_i = \{ \log_{10} I_i(x_i, y_i) - \log_{10} [I_i(x_i, y_i) \times \mathcal{F}(x_i, y_i)] \} \quad (4.3)$$

From the basic definition of the captured image \mathcal{I}_i can be defined as

$$I_i = \mathcal{L}_i(x_i, y_i) \rho(x_i, y_i) \quad (4.4)$$

\mathcal{L}_i is the illumination and ρ is the reflection coefficient. The previous equation can be further written as

$$\mathcal{R}_i = \log_{10} \left\{ \frac{\mathcal{L}_i(x_i, y_i) \rho(x_i, y_i)}{\mathcal{L}_i(x_i, y_i) \rho(x_i, y_i) \times \mathcal{R}_i} \right\} \quad (4.5)$$

The performance of the single Retinex largely depends on the value of σ . Due to the problem in selecting the value of σ , dynamic range of color rendition always needs to be a trade-off. To overcome this issue, Jobson et al. [222] proposed the multi-scale Retinex formula

$$\mathcal{R}_{MSR_i} = \sum_{n=1}^N w_n \mathcal{R}_i \quad (4.6)$$

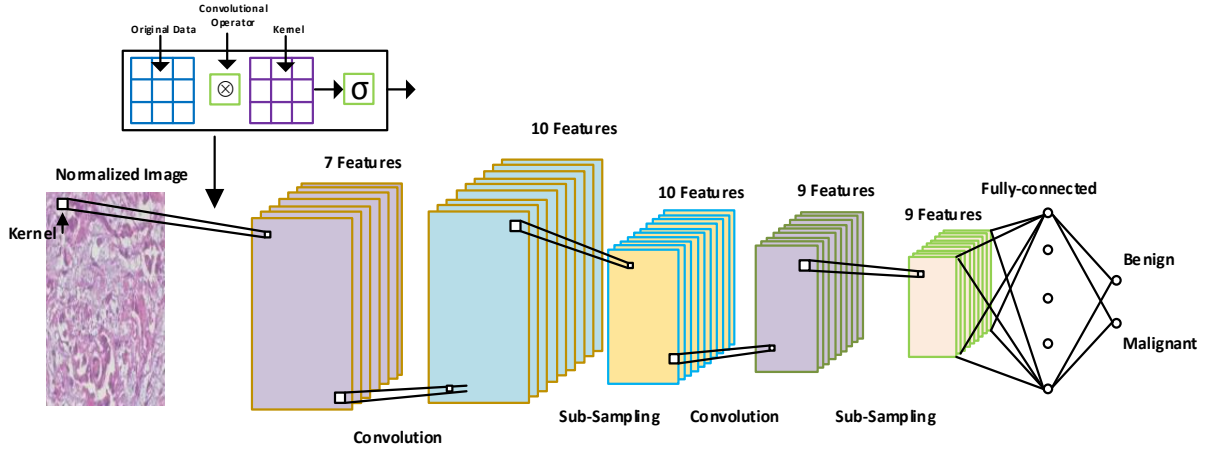


Figure 4.3: Workflow of a Convolutional Neural Network

which satisfies the condition $\sum_{n=1}^N w_n = 1$, where n is the scale and w_n is the weight of that particular scale.

Each of the images has been passed through the Retinex algorithm to improve the local statistical information, and the preprocessed image has been passed through the CNN model for the image classification described in the following section.

4.4 Utilised Convolutional Neural Network

Let, $I_{m \times n}$ represent the images where m and n are the length and width of the image. In a CNN, each image is convolved with the 2-D kernel $W_{p,q}^S$ where p and q represent the kernel size and s represents the used kernel. The output of the convolved signal between $I_{m \times n}$ and $W_{p,q}^S$ is

$$\mathcal{C}_{(x,y)}^S = I_{m \times n} * W_{p,q}^S \quad (4.7)$$

Each entry of $\mathcal{C}_{(x,y)}^S$ can be represented as

$$\mathcal{C}_{(x,y)}^S(i,j) = \sigma \left(\sum_{u=i}^p \sum_{v=j}^q I_{m \times n}(u-i, v-j) * W_{p,q}^S(u,v) + b^s \right) \quad (4.8)$$

where σ is a nonlinear function. For our application we have utilised Rectifier Linear Unit (ReLU) methods which work as follows:

$$\sigma(x) = \max(0, x). \quad (4.9)$$

For the reduction of the classification complexity we have utilised a Sub-sampling operation. At the end of the network a flat layer has been introduced followed by fully connected and Soft-Max layers which allow the network to behave as conventional Neural Network. The basic CNN-based classification model is represented in Figure 7.7.

The layer before the softmax layer can be represented as

$$\mathcal{H}_g^{\text{end}} = \sigma(w^{\text{end}} * \mathcal{H}_g^{\text{end}-1} + b^{\text{end}}). \quad (4.10)$$

As we are working on a binary classification, the soft-max regression output can be represented as

$$\bar{y}_g = \frac{\exp(\mathcal{H}_g^{\text{end}})}{\sum_{g=1}^2 \exp(\mathcal{H}_g^{\text{end}})} \quad (4.11)$$

The predicted class \bar{p} will be

$$\begin{aligned} \bar{g} &= \arg \max_g \bar{y}_p \\ &= \arg \max_g \bar{h}_p^{\text{end}} \end{aligned} \quad (4.12)$$

Using the basic properties of the CNN model we have utilised following three model for the image classification as:

Model-1 (Conventional Model)

In this model we have utilised four convolutional layers C-1 to C-4 consecutively with 5 by 5 kernels. Each layer includes 16 feature maps and an ReLU rectifier unit. A maximum pooling layer has been placed between layers C-1 and C-2, C-2 and C-3, C3 and C-4 and named MP-1, MP-2, MP-3 and MP-4. After the C-4 layer we have utilised a Flat Layer, Drop-out (0.10%) layer and a Soft-Max layer consecutively to classify Benign and Malignant images.

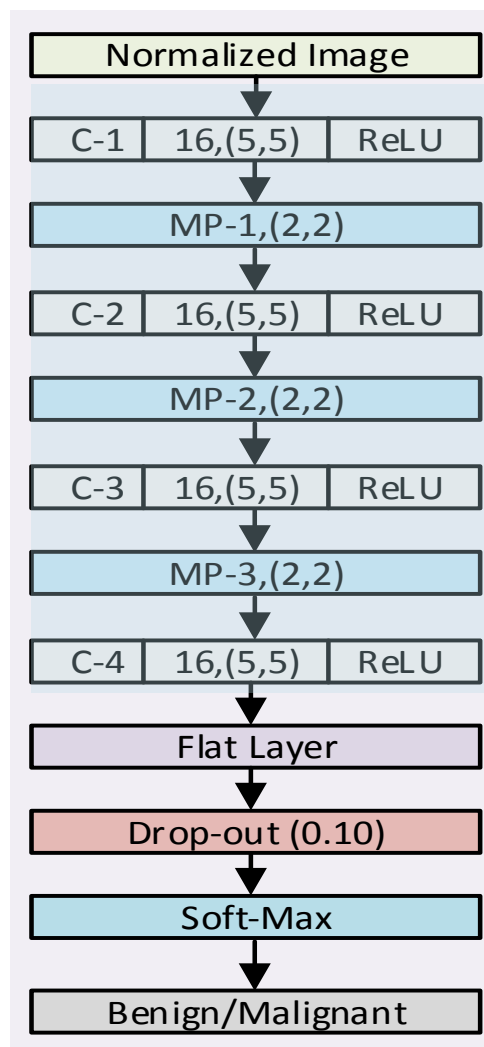


Figure 4.4: Conventional Model

Model-2 (Merge Model)

Adding more layers in a CNN will not always improve the system performance. Experiment has found that increasing the number of network layers may cause saturation of the accuracy and performance. To overcome this issue, Kaiming He [84] proposed residual learning methods which dramatically improve the system performance. In their original

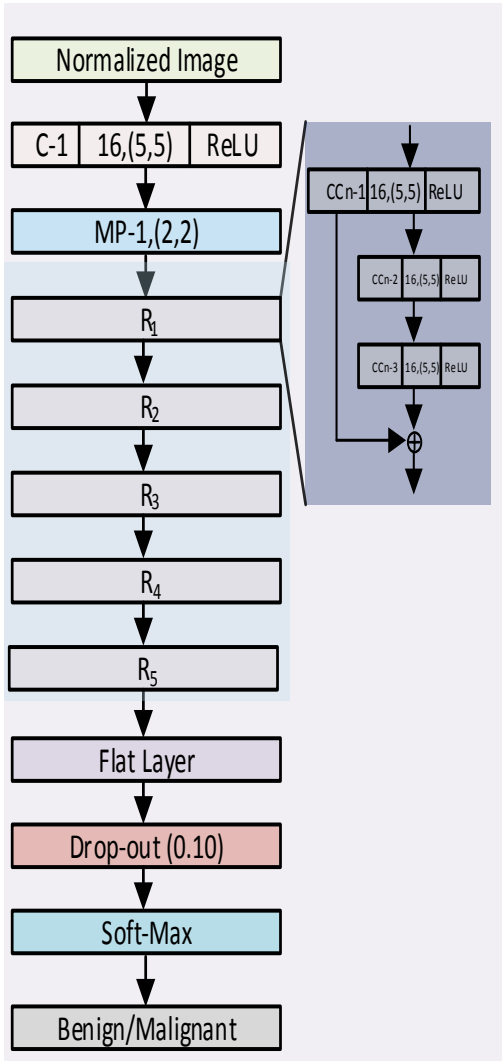


Figure 4.5: CNN model with Residual Block

model, identity mapping is added to the stacked-layers output. This layer does not in-

crease the complexity and it also satisfies end-to-end training with back-propagation. It is believed that, due to the reference identity, optimizing the residual network is easier than for the original network. The first step of the model performs a 2D convolutional operation (C-1) using a 5 by 5 kernel, and produces 16 feature maps. After the C-1 layer, the Max-Pooling (MP-1) operation is performed. A second 2-D convolution operation (C-2) is performed following a Max-Pooling (MP-2) operation. This C-2 layer also utilises a 5 by 5 kernel and produces 16 feature maps. After the C-2 layer, a few Residual ($i = n$) blocks are utilised, where the value of $n = 1$ to 5. The first residual operation can be represented as

$$X_{C-2}^1 = \sigma[F_1^1(X_{C-2}, W_{11})] \quad (4.13)$$

$$R_1 = \sigma[F_1^2(X_{C-2}^1, W_{12}) + X_{C-2}] \quad (4.14)$$

where X_{C-2} represents the output of the convolutional layer C-2. The value of R_1 is used to calculate the value of R_2 and the same procedure continues. The output of the residual layer R_5 is passed through a Max-Pooling operator. After that a Flat layer, a Drop-out layer and Soft-Max layer are placed consecutively to get the Benign and Malignant classified output.

Model-3 (MaxMin Convolutional Model)

The working principle of the ReLU rectifier shows that it gives identical behavior for all negative values. Small and large negative values clearly show different behavior. To utilise strong negative-value information in the CNN network, M. Blot proposed the MaxMin convolutional Neural Network [223].

Let h be a filter and $h^- = -h$ represent the negative filter. If we apply the convolutional operation on the signal x then the overall performance can be written as

$$x \star h^- = x \star -h = -x \star h \quad (4.15)$$

Instead of using the negative filter, in MaxMin theory the original signal is duplicated along with the negative signal, finally utilizing both the original and duplicate signals for the further processing steps.

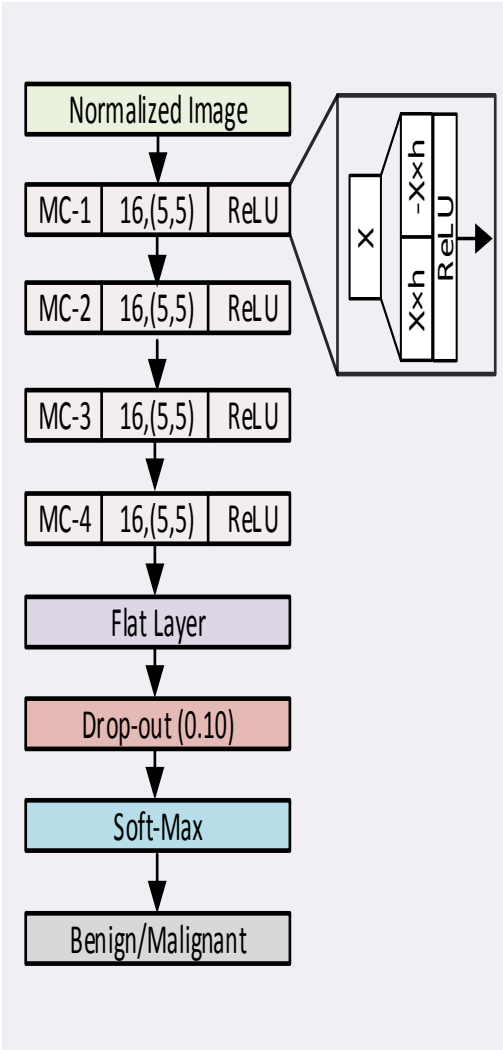


Figure 4.6: Max-Min Model

We have utilised four consecutive MaxMin convolutions MC-1 to MC-4. Each of the two MaxMin layers produces sixteen feature maps. After the MC-4 layer a Max-Pooling, Flat layer, Drop-out layer and a Soft-Max layer have been placed one after another for the Benign and Malignant classification.

4.5 Results and Discussion

Table 4.1: Performances of the different Cases on different datasets

	Case	Accuracy (%)	Specificity (%)	FPR (%)	FNR (%)	Recall (%)	Precision (%)	F-Measure (%)
40×	Model-1	85.00	73.36	26.63	10.36	89.63	88.00	89.00
	Model-2	80.63	64.73	35.32	12.28	87.71	85.00	86.00
	Model-3	80.13	75.00	25.00	17.59	82.40	88.00	85.00
100×	Model-1	85.36	70.28	29.14	08.63	91.36	89.00	90.00
	Model-2	80.65	66.28	33.71	13.63	86.36	87.00	86.00
	Model-3	84.60	61.14	38.85	6.80	93.10	86.00	89.00
200×	Model-1	85.28	70.20	29.79	03.71	92.60	86.00	89.00
	Model-2	76.19	30.30	69.00	07.40	98.52	74.00	85.00
	Model-3	85.45	71.71	28.28	07.86	92.13	87.00	89.00
400×	Model-1	85.16	68.88	31.14	06.60	93.38	86.00	89.00
	Model-2	83.15	69.39	30.60	9.09	90.08	85.00	88.00
	Model-3	80.13	67.75	32.24	9.36	90.63	85.00	88.00

Our classification experiment has been performed on the BreakHis breast-image dataset [224], which contains four sets of images "m×" here $m = \{40, 100, 200, 400\}$ and \times represents the magnification factor.

Table 4.1 represents the Accuracy, Specificity, False Positive Rate (FPR), False Negative Rate (FNR), Recall, Precision and F-Measure values for the experiment, where Benign has been considered as the Negative class and Malignant has been considered as the Positive Class. The first quarter of the table shows the performance of the three models on the 40× dataset. For the Accuracy, the best performance is achieved when Model-1 is utilised, at 85.00%. Model-2 and Model-3 give almost the same performance, at around 80.50%. The Recall values are 89.63%, 87.71% and 82.40% for Model-1, Model-2 and Model-3 respectively. This indicates that almost 10% of the Malignant data is mis-classified as Benign data when we utilise the Model-1 algorithm on the 40× dataset. The worst Recall value was given by Model-3; in this particular case almost 18% of Malignant data has been mis-classified as Benign data. Interestingly, for the specificity criterion, Model-3

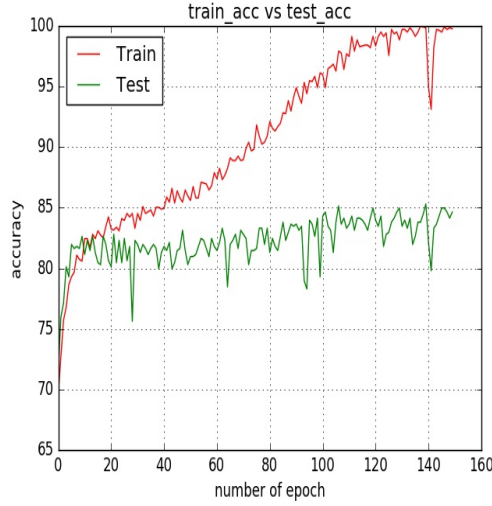
shows better performance than Model-1 and Model-2, at exactly 75.00%. For the specificity case, Model-1 provides 73.36%. The precision values for Model-1 and Model-3 are almost equal at 88.00%. For the F-Measure performance Model-1 provides the best value at 89.00%.

For the 100 \times dataset, the Accuracy values for Model-1 and Model-3 differ marginally, at exactly 85.36% and 84.60% respectively. In this particular dataset Model-3 provides the worst performance at exactly 80.65%. The best Recall value is 93.10% which is achieved by Model-3, whereas Model-1 provides 91.36% recall values. However for the specificity Model-3 provides the worst performance at 61.14%; this indicates that almost 40.00% of Benign data has been mis-classified as Malignant data. In the Specificity criterion, Model-1 provides the best performance at 70.28%. The best Precision and Recall values are achieved when we utilise Model-1, at exactly 89.00% and 90.00% respectively.

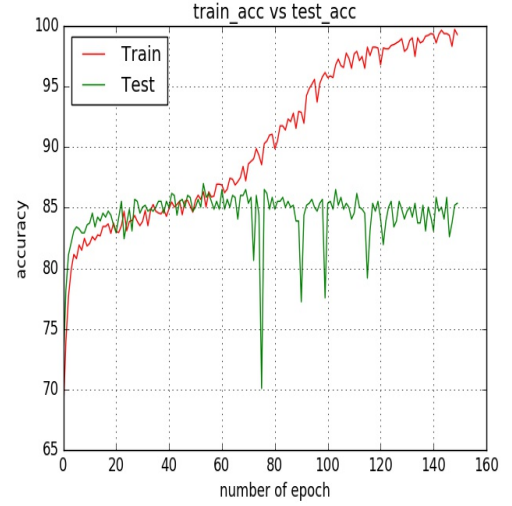
For the 200 \times dataset the best Accuracy is achieved when we utilise Model-3, at 85.45%, which is slightly better than for Model-1 at 85.28%. In this particular dataset Model-2 provides somewhat worse Accuracy performance at 76.19%. For the criteria Specificity, FPR, FNR, Recall, Precision and F-Measure the values are almost the same for Model-1 and Model-2. Interestingly, though the overall Accuracy performance of Model-2 is poor, it shows the best Recall value, 98.52%. This indicates that it classifies almost all the Malignant data as Malignant. However for Model-2, the Specificity value is very poor at 30.30%, which indicates that it mis-classifies 70.00% of the Benign data as Malignant, so that Model-2 is very poorly Specific and Highly sensitive on the 200 \times dataset.

For the 400 \times dataset the best Accuracy value is achieved when we utilise Model-1, at 85.16%, while the Model-2 provides the second-best Accuracy performance, at 83.15%. The best Specificity value is provided by Model-2 at 69.39%, this indicating that almost 30.00% of the Benign data has been mis-classified as Malignant data. For the Recall case Model-1 provides the best performance at 93.38%. The best Precision and Recall values

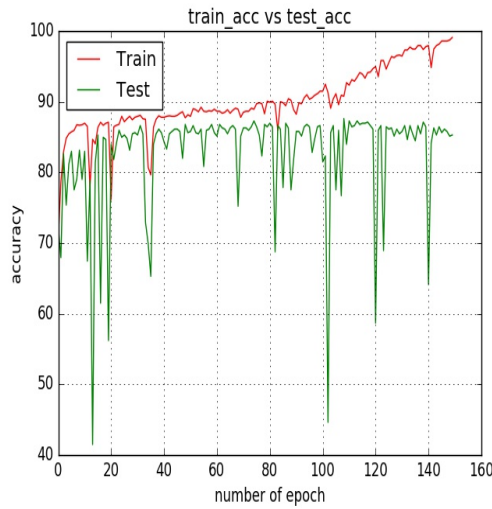
are 86.00 % and 89.00% which is achieved when we utilise Model-1. Overall, Model-1 provides the best Accuracy performance for all the datasets except the $200\times$ dataset.



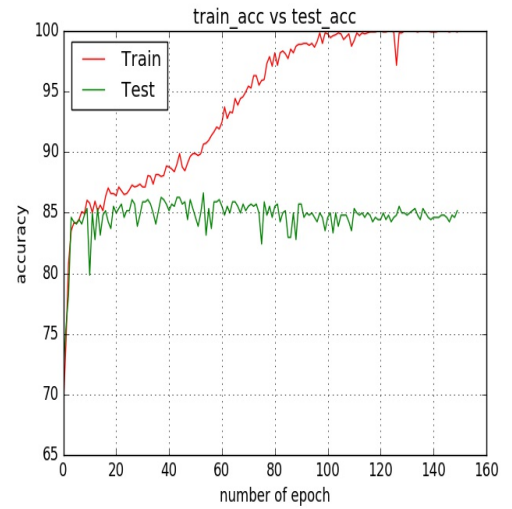
(a)



(b)



(c)



(d)

Figure 4.7: (a), (b), (c) and (d) represent the training and test accuracy comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.

Figure 5.6 shows the accuracy performance of the training and test datasets up-to epoch 150 for Model-1 for all the datasets. When we use Model-1 the best performance is

achieved for the $100\times$ dataset as seen in Figure 5.6 (b). This figure also shows that, up to around epoch 20, the Train Accuracy shows better performance than the Test Accuracy. From epoch 20 to around 55 both the Train and Test accuracy are almost the same. After around the first 55 epochs, the Train Accuracy outperforms the Test Accuracy, and this continues as the epoch increases. Interestingly the performance of the Test Accuracy remains almost constant after around epoch 55. When we use Model-1 almost all the datasets provide the same performance, however Figure 5.6 (c) shows that up to around epoch 20 the Train and Test accuracy remain almost constant and after epoch 100 the Train Accuracy outperforms the Test Accuracy.

Figure 4.8 (a), (b), (c) and (d) show the loss performance for Model-1. Interestingly, when we utilise Model-1 and the $40\times$, $100\times$ and $400\times$ datasets, after some initial epochs the Train loss and Test loss show a large divergence, and this divergency goes higher as the epoch goes higher. When we utilise the $100\times$ dataset, up to around epoch 60 the Train and Test losses remain almost the same. After around epoch 60 the Train loss decreases, however the test loss remains almost constant. This reflects the justification of the accuracy performance of the database $100\times$ as presented in Figure 5.6 (b), which shows that the Train and Test loss remain almost the same when we utilise the $200\times$ dataset. After around epoch 100 the Train and Test loss differences increase rapidly.

Figure 4.9 (a), (b), (c) and (d) show the Matthews Correlation Coefficient (MCC) performance for Model-1. These values lie in the range -1 to +1, where +1 indicates the best performance. For Model-1, all the values of the Train MCC remain constant around 0.70 whereas as the epoch advances the Test MCC values reach 1.00.

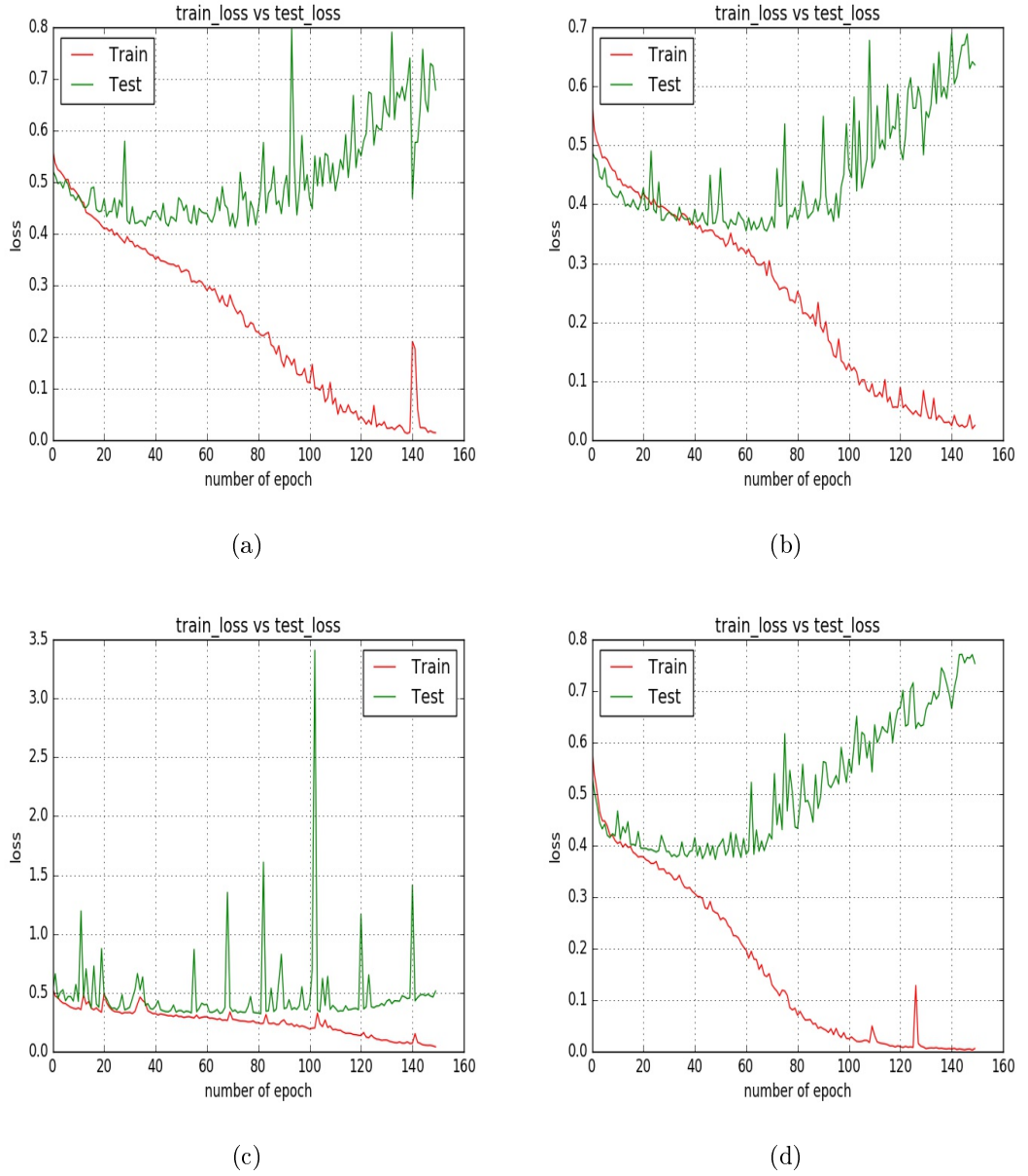
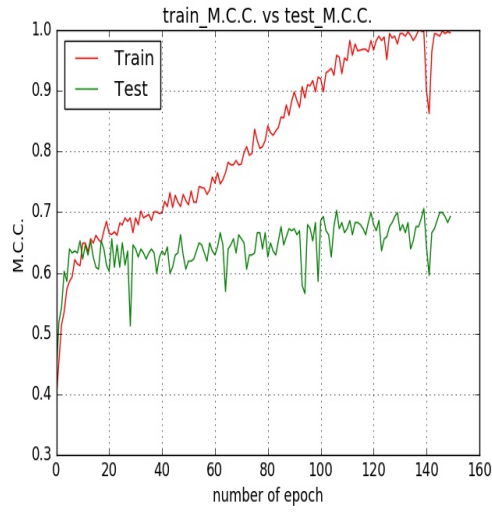


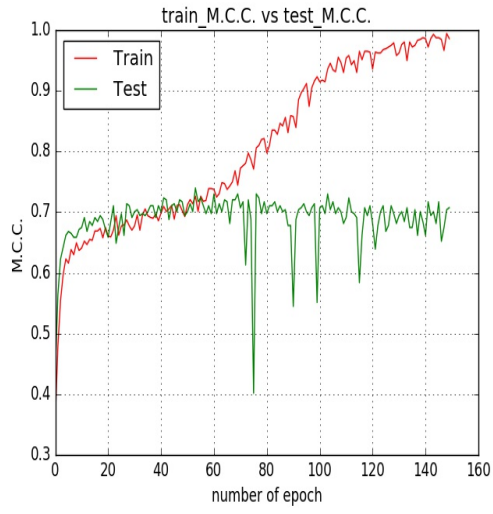
Figure 4.8: (a), (b), (c) and (d) represent the training and test loss comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.

4.5.1 Time and Parameters required

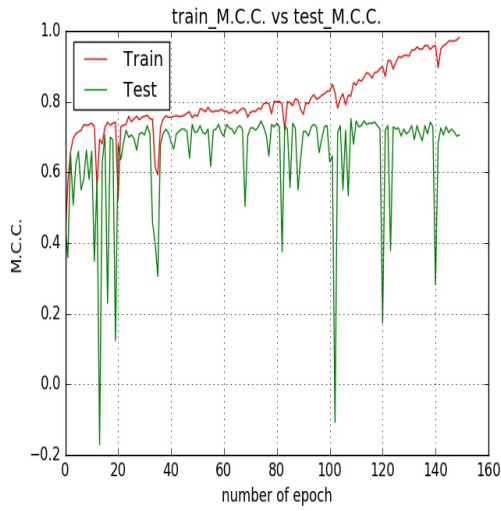
Table 4.2 shows the number of required parameters and the time required to run per epoch for the three models. Model-1 requires less time and fewer parameters to perform the classification. Model 2 requires the most parameters (105650) to run the whole classification



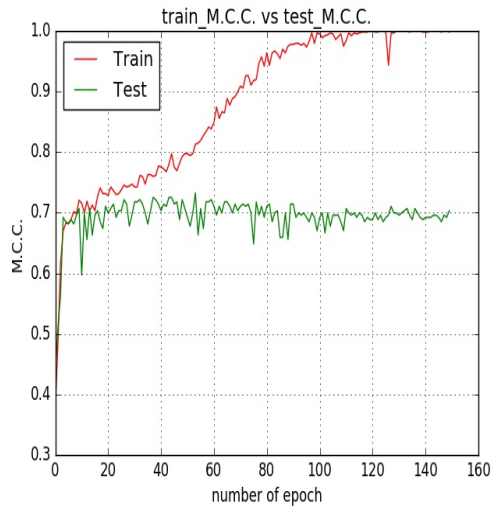
(a)



(b)



(c)



(d)

Figure 4.9: (a), (b), (c) and (d) represent the training and test M.C.C comparison for Model-1 on the $40\times$, $100\times$, $200\times$, $400\times$ datasets.

performance.

Table 4.2: Time (per/epoch) and parameters required to run the models

	Required Parameters	Time/Epoch (s)
Model-1	27122	6
Model-2	105650	14
Model-3	41714	19

4.6 Conclusion

In this chapter we have classified histopathological breast images (from the BreakHis Dataset) into Benign and Malignant classes using a Convolutional Neural Network technique. For illumination correction and image enhancement we have utilised the Retinex algorithm. We have utilised Conventional, Residual and MaxMin Convolutional Neural Networks, with the best results achieved when we utilised the Conventional model. Most of the recent work on this particular dataset has provided only Accuracy information. However in this chapter, along with the Accuracy measure, we have provided detailed information about the Specificity, FPR, FNR, Recall, Precision and F-Measure values. Computational complexity and time is a big issue in the CNN-model based system analysis. In this chapter we have provided the information about the required number of parameters and time to perform the operation, which provide some demonstration about the complexity of the model.

Chapter 5

Local and Global Feature Utilisation for Breast-Image Classification by Convolutional Neural Network

5.1 Abstract

Convolutional Neural Networks (CNN) have brought a revolutionary improvement to image analysis, specially in the image classification field. The technique of natural image classification using the CNN method has been deliberately utilised for medical image classification with some advanced engineering. However, so far in most of the cases CNN model classifies images based on global features extraction from the raw images. In this chapter we have utilised both raw images and some hand-crafted features, and later we classify images using a CNN network. For the classification purposes we have utilised the BreakHis dataset and achieved a 96.00% accuracy, which is a state-of-the-art result on

Published as: A. A. Nahid and Y. Kong, “Local and Global Feature Utilization for Breast Image Classification by Convolutional Neural Network”, in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, pp. 1-6, Nov. 2017

this dataset.

5.2 Introduction

All the cells of the body maintain a cyclic order, where new cells grow as replacements of old cells. However this might not be always the case. Sometimes few of the cells show some abnormality, and these cells can continuously grow and create a cancer. This disease can occur in any part of the body and later spread to any other part. Different cancer diseases such as liver, skin, breast etc. are more dominant. However, of all the available cancer diseases, women are more vulnerable to breast cancer due to their physical anatomy. As an example, the following graph shows the last 12 years of statistics about womens' deaths due to cancer in Australia (population 20-25 million). Though this graph is an example, it might be considered as representative of the current condition of the whole world.

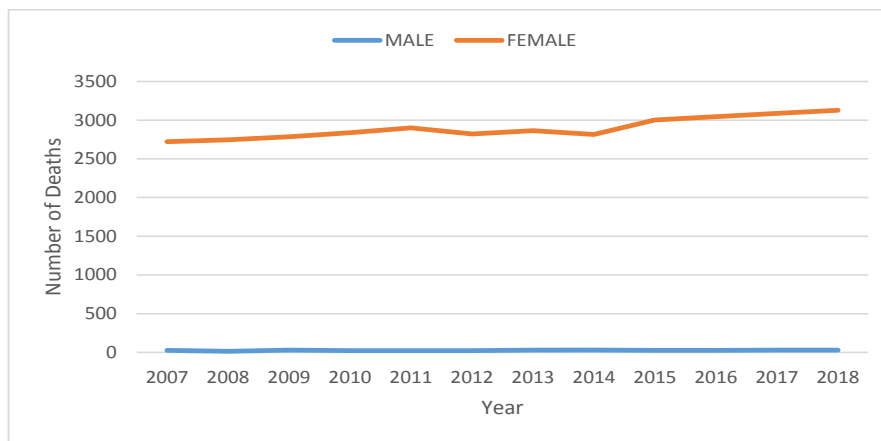


Figure 5.1: Male and Female cancer death statistics for the last decade in Australia.

Cancer-suffering people are always in a vulnerable condition. Proper identification of the cancer can save or at least reduce the miserable condition of cancer-affected people. Among other techniques, inspection of biomedical images is an important method for the analysis of cancer. Different biomedical imaging techniques have been available

such as MRI, X-Ray, histopathological etc. Among the digital photographic techniques, histopathological images are very popular with physicians and doctors for cancer inspections. However inspection of these histopathological images is always time-consuming and also requires an extra level of expertise. Modern digital image analysis techniques allow automatic detection as well as classification of the images, which help doctors and physicians to get a second opinion about the cancer.

Different methods and techniques have been utilised for the image classification. However the state-of-the-art image classification from Convolutional Neural Networks (CNN) has brought a revolutionary change in biomedical image classification. The history of the use of CNN for biomedical image classification is a long one. "Neocognitron" is the first CNN model, proposed by K. Fukushima et al. for the recognition of stimulus patterns [79]. The CNN model was first utilised by Y. Wu et al. for mammogram image classification [80]. After a long break, the CNN method again got momentum after the work of A. Krizhevsky with his model AlexNet [219]. After the A. Krizhevsky's model, few other models have been introduced, like Visual Geometry Group (VGG-16, VGG-19), GoogleNet, which are actually advanced-engineering versions of the AlexNet model.

B. Sahiner et al. [81] and J. Arevalo et al. [93] classified mammogram images utilizing global features and achieved Receiver Operating Characteristics (ROC) values of 0.87 and 0.826, respectively. F. Spanhol et al. [7] classified histopathology images using a CNN method, and the best accuracy achieved was 89.00%. J. Xu et al. [225] also classify histopathology images where a best ROC 0.93 is achieved using DCNN-Ncut-SVM.

Most of the classification using CNN utilised global feature extraction from the raw images. However H. Rezaeilouyeh et al. [226] classify histopathological images using CNN, where they utilised local as well as global features. They have utilised a Shearlet transform for extracting local features. They obtained the best accuracy 86 ± 3.00 % when they utilised "RGB+ magnitude+phase of Shearlets" together. K. Sharma et al. utilised

the Gray Level Co-occurrence Matrix (GLCM) and the Gray Level Difference Method (GLDM) together for the Mammogram image classification. The best accuracy achieved was 75.23% and 72.34%, for the fatty and dense tissue classifications respectively.

Inspired by the recent advancement of the CNN model, in this chapter we have classified a histopathological image (BreakHis) database using CNN where we have utilised both global and local features together. Section 5.2 gives introductory literature and the current status of breast image classification using CNN, Section 5.3 describes the overall architecture along with the feature extraction techniques used on the images, Section 5.4 describes the working principle of CNN along with the model which we have utilised for the classification purposes, Section 5.5 describes the results and explains some performance measuring parameters, and lastly we conclude our chapter at Section 5.6.

5.3 Overall Architecture for Classification

A conventional CNN extracts global feature information from the raw images and performs further post-processing operations. However some work has recently been performed on image classification utilizing hand-crafted local features such as GLCM, GLDM along with the CNN model for the image classification. To classify the data, we have utilised both the raw images and hand-crafted information (Histogram information and Local Binary Pattern (LBP)) together, for the image classification. The overall classification model is presented in Figure 5.2. We have named our algorithm "C-H" where raw images along with the histogram information are used, however when we utilised raw images and the LBP information together we named this algorithm "C-L".

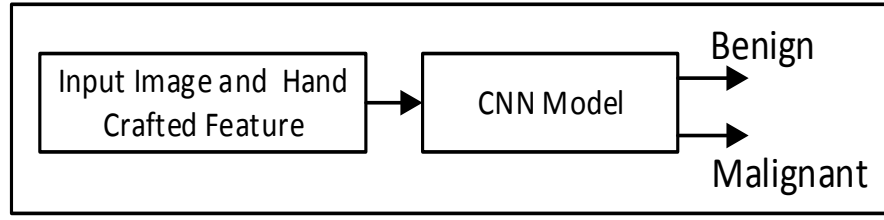


Figure 5.2: Overall image-classification model

5.3.1 Handcrafted Features

Among different hand-crafted feature selection methods we have selected the following two features for classification:

Histogram Information

Combination of the Red, Green, and Blue light information creates an RGB image $I(x, y)^{RGB}$. As RGB is additive in nature, we can represent this RGB image as

$$I(x, y)^{RGB} = I(x, y)^R + I(x, y)^G + I(x, y)^B$$

A graphical display which represents the frequency of each of the particular intensities in an image is known as a histogram. Let $H(x, y)_L^{RGB}$, $H(x, y)_L^R$, $H(x, y)_L^G$, $H(x, y)_L^B$ represent the histogram of $I(x, y)^{RGB}$, $I(x, y)^R$, $I(x, y)^G$, $I(x, y)^B$, where L is the level number, in this case $L = 0, \dots, 255$.

Local Binary Pattern

The Local Binary Pattern (LBP) represents an image $I(x, y)$ by a matrix which contains integer labels, as proposed by Ojala et al. [25], [26]. Lately histograms of this matrix have been used for further image analysis.

Let the strength g_c of any arbitrary pixel position (x_c, y_c) of the image $I(x, y)$ be

represented as

$$g_c = I(x_c, y_c) \quad (5.1)$$

and consider a circle of radius R with evenly distributed p points (x_p, y_p) around the central pixel (x_c, y_c) . The strength of the point (x_p, y_p) can be represented as

$$g_p = I(x_p, y_p) \quad (5.2)$$

where

$$y_p = y_c - R \sin\left(\frac{2\pi p}{P}\right) \quad (5.3)$$

and

$$x_p = x_c + R \cos\left(\frac{2\pi p}{P}\right). \quad (5.4)$$

Then the LBP can be defined as

$$\text{LBP}_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (5.5)$$

where

$$s(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{otherwise} . \end{cases}$$

5.4 Convolutional Neural Network and our proposed model

A CNN model is the combination of a few intermediate mathematical structures which create different layers. Among all the other layers, the convolutional layer is considered as the most important part for a CNN model and can be considered as the backbone of the model. A two-dimensional weight matrix, named a kernel, of size $m \times n$ is scanned through the input data for the convolutional operation. This ensures the local connectivity

and weight-sharing property. The number of steps a kernel will move through the image is known as the stride. To overcome the edge effect of the images, a method known as zero padding has been utilised. The output of each of the kernel operations is passed through a rectifier functions such as Rectified Linear Unit (ReLU), Leaky ReLU, TanH, Sigmoid etc. The Sigmoid function can be defined as

$$\sigma(x) = \frac{1}{(1 + \exp^{-x})}. \quad (5.6)$$

However the most effective rectifier is ReLU. The ReLU method converts all the information into zero if it is less than or equal to zero, and passes all the other data as is. Figure 5.4 shows the working principle of the ReLU method:

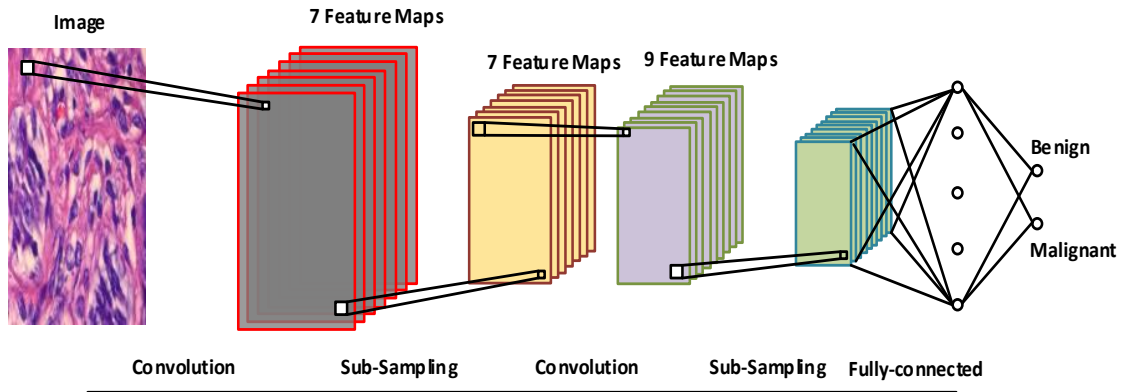


Figure 5.3: Workflow of a Convolutional Neural Network

$$\sigma(x) = \max(0, x). \quad (5.7)$$

Sub-sampling is the procedure of reducing the dimensionality of each of the feature maps of a particular layer; this kind of operation is also known as pooling. Actually, it reduces the amount of feature information from the the overall data and hence, reduces the overall computational complexity of the model as well. To do this $s \times s$ patch units are utilised. The two most popular pooling methods are

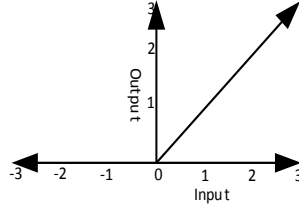


Figure 5.4: ReLU Opearation

- Max-Pooling
- Average Pooling.

All the neurons of the most immediate layer of a fully connected layer are completely connected with the fully connected layer, like a conventional Neural Network. Let f_j^{l-1} represent the j^{th} feature map at the layer $l - 1$. The j^{th} feature map at the layer l can be represented as

$$f_j^l = \sigma\left(\sum_{i=1}^{N^{l-1}} f_i^{l-1} * k_{i,j} + b_j^l\right) \quad (5.8)$$

where N^{l-1} represents the number of feature maps at the $l - 1^{\text{th}}$ layer, $k_{i,j}$ represents the kernel function and b_j^l represents the bias at l , where σ performs a nonlinear function operation. For the classification a CNN model utilised a fully connected layer and a Soft-Max layer. The layer before the soft-max layer can be represented as

$$h_p^{\text{end}} = w^{\text{end}} * h_p^{\text{end}-1} + b^{\text{end}} \quad (5.9)$$

As we are working on a binary classification, the Soft-Max regression output can be represented as

$$\bar{y}_p = \frac{\exp(h_p^{\text{end}})}{\sum_{p=1}^2 \exp(h_p^{\text{end}})} \quad (5.10)$$

The predicted class \bar{P} will be

$$\begin{aligned} \bar{p} &= \arg \max_p \bar{y}_p \\ &= \arg \max_p \bar{h}_p^{\text{end}} \end{aligned} \quad (5.11)$$

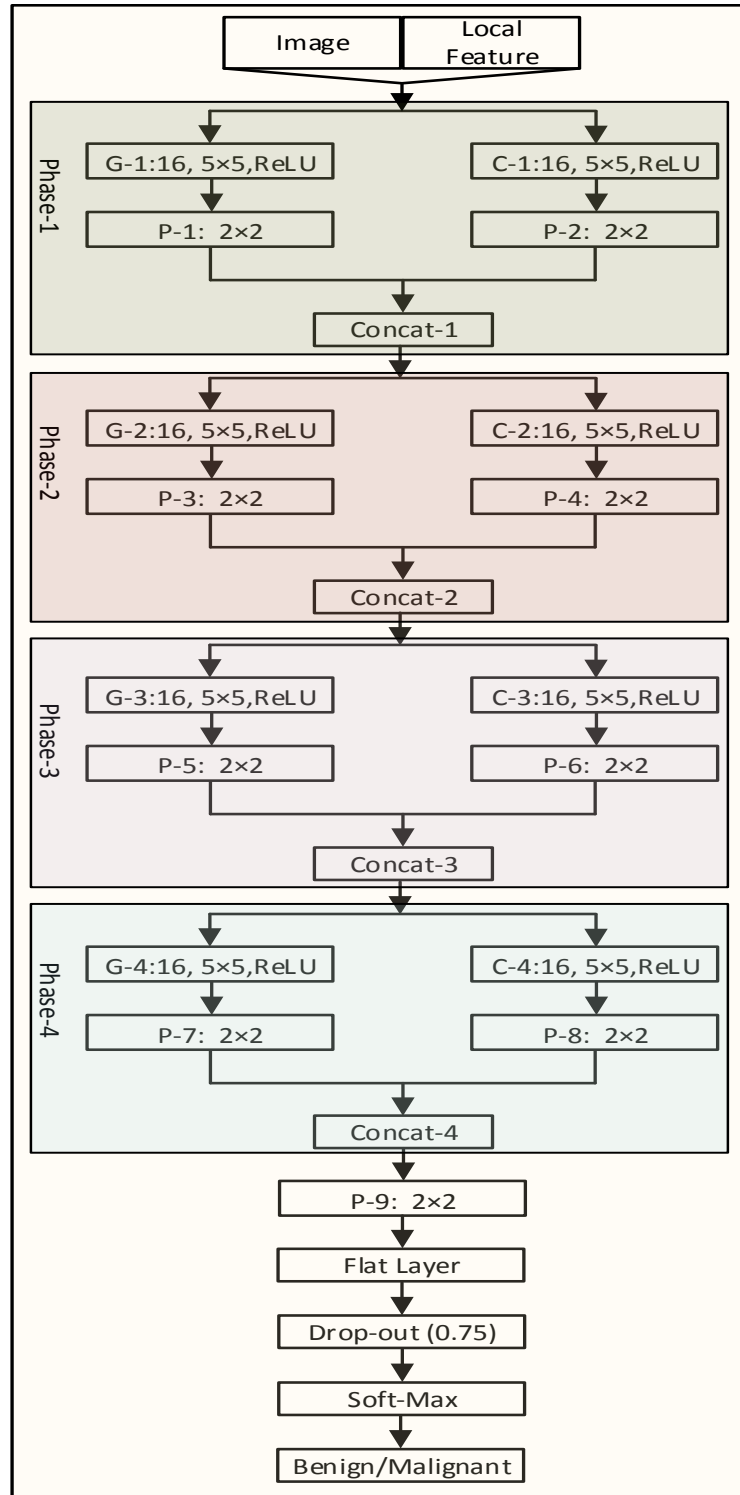


Figure 5.5: The CNN model utilised for the classification

For the breast-image classification we have utilised the following model:

5.4.1 Our Model for Classification

In our model we have divided the convolutional operation into four phases. In Phase-1, convolutional operations are performed on the raw images as well as the local feature information with the two parallel convolutional layers G-1 and C-1. Both the convolutional layers G-1 and C-1 produce 16 feature maps and utilise a 5×5 kernel filter with the ReLU rectifier. The Max-Pooling operation with kernel size 2×2 is performed in layers named P-1 and P-2. The outputs of the P-1 and P-2 layers are concatenated together in the concatenated layer named Concat-1. The same procedure is carried out in Phase-2, Phase-3 and Phase-4 consecutively. The output of the layer named Concat-4 is passed through the Max-Pooling layer named P-9 with kernel size 2×2 . After the Flat-layers we have utilised a Drop-out layer where 75.00% of the data has been dropped. After the Drop-out layer the model uses a Soft-Max layer for the image classification into Benign and Malignant images.

5.5 Results and Discussion

5.5.1 Performance-Measuring Parameters

A Confusion Matrix is a two-dimensional table which is used to give a visual perception of classification experiments [43]. The $(i,j)^{th}$ position of the confusion table indicates the number of times that the i^{th} object is classified as the j^{th} object. The diagonal of this matrix indicates the number of times the objects are correctly classified. Figure 7.9 shows a graphical representation of a Confusion Matrix for the binary classification case. Among the different classification performance properties, this matrix will provide following parameters:

		Hypothesised Class	
		Benign	Malignant
True Class	Benign	True Negative (TN)	False Positive (FP)
	Malignant	False Negative (FN)	True Positive (TP)

Figure 5.6: Confusion Matrix

- Recall is defined as $\text{Recall} = \frac{TP}{TP+FN}$.
- Precision is defined as: $\text{Precision} = \frac{TP}{TP+FP}$.
- Specificity is defined as: $\text{Specificity} = \frac{TN}{TN+FP}$.
- Accuracy is defined as $\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$.
- F-1 score is defined as $F_1 = \frac{2 \times \text{Recall}}{2 \times \text{Recall} + FP + FN}$.
- Matthew Correlation Coefficient (MCC): MCC is a performance parameter of a binary classifier, in the range $\{-1 \text{ to } +1\}$. If the MCC values tend more towards +1, the classifier gives a more accurate result, the opposite condition will occur if the value of the MCC tend towards -1. MCC can be defined as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5.12)$$

5.5.2 Results and Discussion

We have utilised the BreakHis dataset which contains four sets of images " $m \times$ " where $m = \{40, 100, 200, 400\}$ and \times represents the magnification factor. Table 5.1 shows the Specificity, Recall, False Negative Rate (FNR), False Positive Rate (FPR), Precision and

F-1 score values for both the "C-H" and "C-L" algorithms. We have considered the validation performance as the test performance because the "BreakHis" dataset contains only a small number of images. In all the following plots, the Blue plot represent train information and the Green plot represent validation/accuracy information.

Table 5.1: The performance of the C-H and C-L algorithms on the BreakHis dataset

		Specificity/ TNR (%)	FPR (%)	FNR (%)	Recall/ TPR (%)	Precision (%)	F-1Score (%)
40×	C-H	94.25	05.74	0.800	92.00	98.00	95.00
	C-L	82.75	17.20	18.80	81.20	92.00	82.00
100×	C-H	92.00	08.00	02.00	97.00	97.00	97.00
	C-L	64.00	36.00	02.00	97.20	88.00	87.00
200×	C-H	92.00	08.00	01.00	99.00	97.00	97.00
	C-L	71.00	29.00	05.00	95.00	86.00	91.00
400×	C-H	97.18	02.10	07.40	92.56	98.82	94.70
	C-L	74.00	26.00	04.00	96.00	88.00	81.00

For the 40× dataset, when we utilised the C-H algorithm, the specificity value was 94.25% which indicates that this algorithm classifies almost 95.00% of the original benign images as benign, however it classifies 08.00% of the malignant images as benign too. This indicates that this algorithm shows more or less the same amount of specific and sensitive information on this particular dataset. When we use the C-L algorithm on the 40× dataset, both the specificity and recall values are almost the same at around 82.00%; this indicates that around 18.00% percentage of the benign data has been mis-classified as malignant and conversely.

When we utilise the C-H algorithm on the 100× dataset, FPR and FNR are 08.00% and 02.00% respectively. This indicates that this dataset is 92.00% specific when we utilise the C-H algorithm, and maintains a very high sensitivity. However, when we utilise the

C-L algorithm on the $100\times$ dataset, we obtained 97.20% recall values; that means that the malignant images are mostly perfectly classified, however the sensitivity value of this case is 64.00%, which gives very poor performance for the benign to benign image classification.

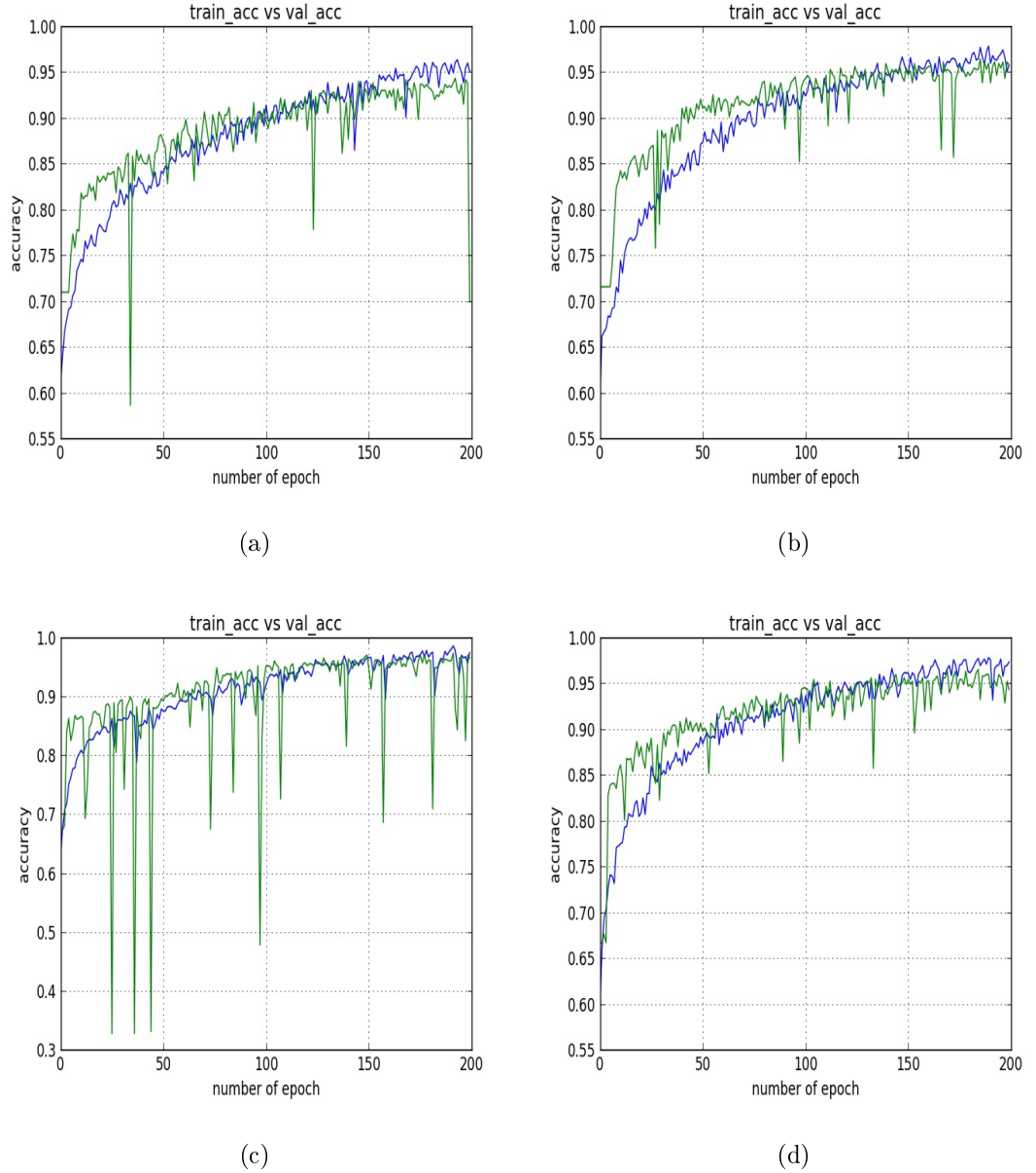


Figure 5.7: Accuracy information when we utilise the C-H algorithm on the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively

When we utilise the C-H algorithm on the $200\times$ dataset, the recall value is 99.00%

which is the best recall value over all the datasets and algorithms. This indicates that this case is very highly sensitive. 8.00% of the benign data has been falsely classified as malignant. When we use the C-L algorithm on this particular dataset, the recall value is 95.00% which indicates that only 5.00% of the malignant data has been mis-classified as benign. This result shows that the C-L algorithm maintains high sensitivity values. However the TNR of this particular case is 29.00% which shows that this case is not very specific.

When we use the C-H algorithm on $400\times$ data set, it shows 97.18% specificity, which indicates that this case has very good performance for benign to benign image classification. However, its recall value is 92.56%, which indicates that around 07.40% of malignant data has been mis-classified as benign. When we utilise the C-L algorithm, the recall value is 96.00% which indicates that around 96.00% of the malignant data has been perfectly classified as malignant, however in this case the FPR value is 26.00% which is quite high.

The above discussion shows that the C-H algorithm gives better performance than the C-L algorithm for all the available datasets. Algorithm C-H on the dataset $400\times$ shows higher specific performance than any other case, and it shows high sensitivity performance on both the $100\times$ and $200\times$ data set.

Figure 5.7 shows the accuracy performance when we utilise the C-H algorithm, where the blue plot represents the train accuracy and the green plot represents the validation (test) accuracy. Up to epoch 100, all the datasets shows better test accuracy performance than the train accuracy performance, which raises the under-fitting issue. Both for the $40\times$ and $200\times$ datasets, after around the epoch 100 the train and validation accuracy remain almost the same. For the $100\times$ dataset from around epoch 100 to 170 the train and test accuracy performance remain almost the same, but after that the train accuracy shows slightly better performance than the validation accuracy. For the $400\times$ dataset from around epoch 100 to 150 the train and test accuracy performance remain almost the

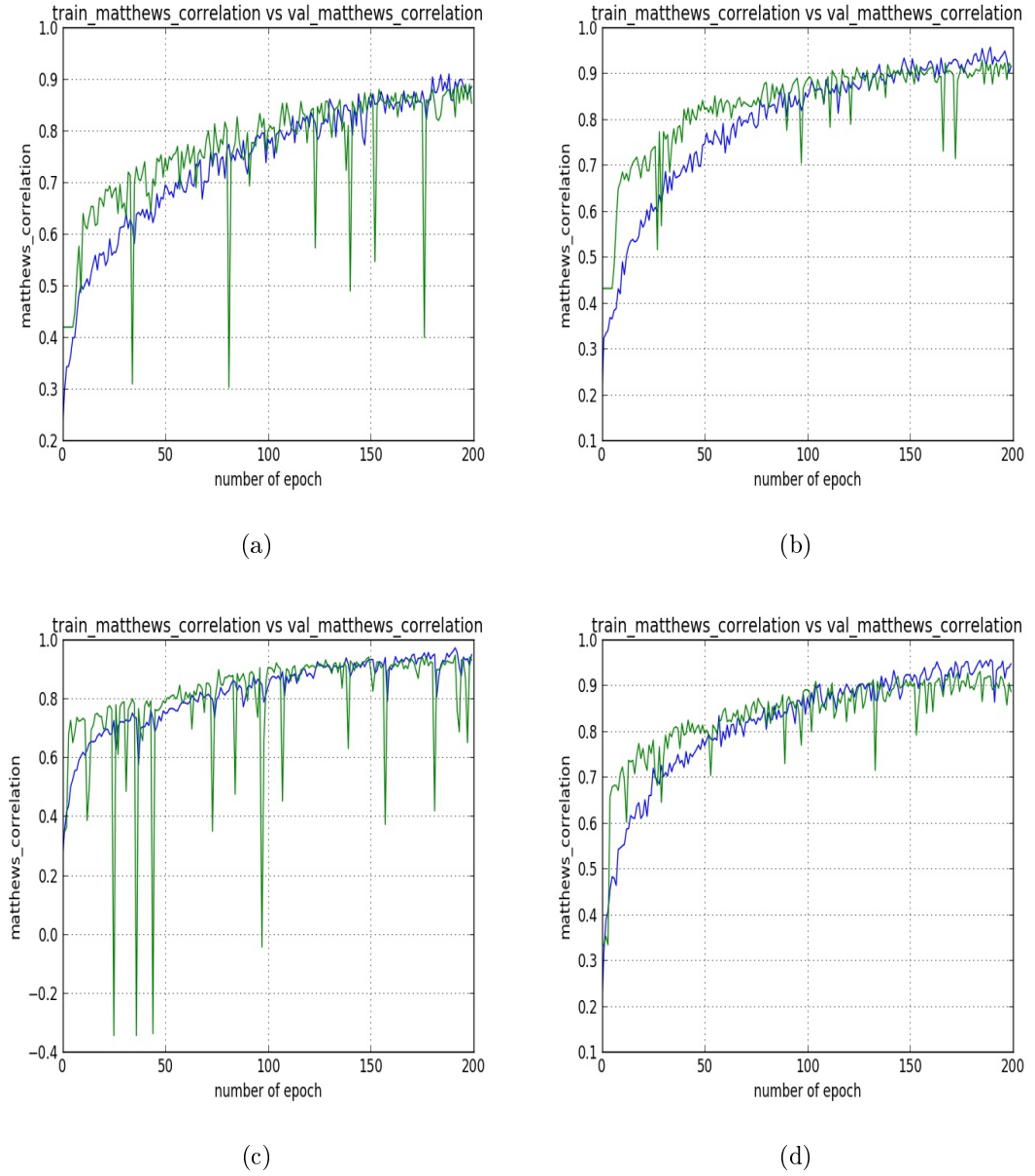


Figure 5.8: M.C.C. information when we utilise the C-H algorithm on the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively

same, but after that the model shows over-fitting performance.

Figure 5.8 a, b, c and d shows the MCC values for the C-H algorithms for the $40\times$, $100\times$, $200\times$ and $400\times$ datasets respectively. Figure 5.8 a, b, c and d show that, initially, both the train and validation (test) MCC values are quite low when we utilised the $40\times$,

100 \times , and 400 \times datasets and as the epoch increases both the train and validation (test) MCC values increase with the epoch and exceeds 00.90. For the 200 \times dataset, with the epoch less than 50, sometimes the MCC values of both the train and validation(test) shows some negative values however around epoch 200 both the train and the validation (test) MCC values show 00.90.

5.6 Conclusion

In this chapter we have classified a breast-cancer (BreakHis) dataset using the state-of-the-art CNN method. Instead of using raw images, we have applied raw images along with some hand-crafted information. The best specificity of 97.18% is achieved when we utilised raw images along with the histogram information on the 400 \times dataset, on the other hand the best sensitivity is achieved when we employed raw images along with LBP information on the 100 \times data set. However the best overall performance is achieved when we utilised raw images along with histogram information on the 100 \times and 200 \times datasets.

Chapter 6

Frequency-Domain Information along with LSTM and GRU Methods for Histopathological Breast-Image Classification

6.1 Abstract

Biomedical image classification has always been a challenging and critical task which has the highest level of importance. The Deep Neural Network (DNN) has been recently introduced for normal image classification and lately introduced for Biomedical image classification with some advanced engineering. In this chapter we have classified an image dataset with a DNN utilizing Long Short Term Memory (LSTM) as well as Gated Recur-

Published as: A. A. Nahid, M. A. Mehrabi and Y. Kong “Frequency-Domain Information Along with LSTM and GRU Methods for Histopathological Breast-Image Classification”, *in 2017 IEEE International Symposium on Signal Processing and Information Technology(ISSPIT)*, IEEE, pp. 1-6, Dec. 2017.

rent Unit (GRU) for breast-image classification. Instead of directly using raw images, we have utilised frequency-domain information for the image classification. Using our model we have obtained 93.01% Accuracy, 94.00% Recall and 94.00% Precision, which is the best available result on this dataset.

6.2 Introduction

Abnormal and unwanted growth of cells is known as cancer and is a serious threat to people throughout the world. At present millions of people are suffering from cancer. As the human body is a combination of cells, and cancer is created from the cells, so cancer can be created in any part of the body and then may migrate to any other part of the body. Among all the other types of cancer, women are more vulnerable to breast cancer than men, due to the anatomical structure of women. Women who are suffering due to breast cancer have a measurable condition that sometimes leads to death. The death of women due to cancer has increased day by day at an alarming rate.

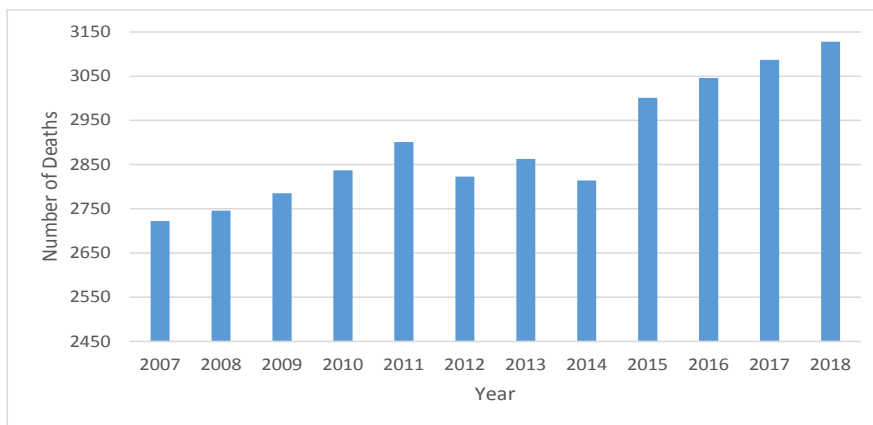


Figure 6.1: Female deaths from breast cancer: statistics for the last decade in Australia.

Figure 6.1 shows the number of women dying from breast cancer in Australia over the last decade. With some exceptions the number of womens' deaths due to breast cancer has

increased every year, and the situation has become worse day by day. Though this shows the women breast-cancer scenario for Australia, this can be considered as representing the whole world scenario.

Treatment of cancer largely depends on identification of the cancer. Inspection of biomedical images is an existing technique for the investigation of cancer. A few biomedical photographic techniques such as Magnetic Resonance Imaging (MRI), Mammogram, Histopathological images are available. Doctors and physicians investigate these images to identify the present status of cancer. Among these biomedical imaging techniques, Histopathological images have been gathered from biopsy of the tissues. Histopathological image investigation always requires expert knowledge to give any opinion about the images, and it requires extra time for the image investigation. However modern digital image-processing techniques have been largely used for biomedical image classification lately. Biomedical image analysis and classification techniques provide a second opinion about the current situation to both the doctors and patients.

Different methods and techniques such as Supervised, Semi-supervised and Un-supervised techniques have been utilised for the image classification. Among the different image-classification techniques, the state-of-the-art Deep Neural Network (DNN) has been introduced for normal image analysis as well as Biomedical image analysis and classification. However the CNN model "neocognitron" which has been introduced by K.Fukushima for biomedical image analysis [79] is considered to be the pioneer work for biomedical image analysis. After the work of K. Fukushima, Y Wu et al. [80] and B. Sahiner et al. [81] have utilised a CNN model for biomedical image analysis from 1990 to 2000. After a long rested, the DNN model again gained interest after the work of A. Krizhevsky et al. [82] in 2012 which introduced the AlexNet model. Basically Alexnet introduced a CNN model for image classification, and lately advanced engineering of the Alexnet model has been introduced for biomedical image classification.

Another branch of the DNN is the Recurrent Neural Network (RNN). The RNN model has been largely utilised for time-series data prediction and classification. The RNN model suffered from the long memory problem and to overcome this issue the Long Short Term Memory (LSTM) [209] and Gated Recurrent Unit (GRU) [227] models have been introduced. This model has mainly classified time-series data. In the conventional CNN model the input data have been considered as high-dimensional vectors. However each data point or a combination of data points of a data matrix can be considered as a data sequence [213]. We have considered that the extracted frequency information maintains a time-series relationship, as adjacent pixels of the images contain similar information [213], and based on this we have utilised LSTM as well as GRU models for Biomedical breast-image classification.

The rest of the chapter is organized as: Section 6.3 will give a brief description of the overall model and frequency-domain feature-extraction techniques as well as the data preparation techniques for the classifier, Section 6.4 will describe the LSTM and GRU models' working principles in brief and also describe our model for breast-cancer image classification, Section 6.5 describes detailed results from our experiment and Section 6.6 concludes our work.

6.3 Overall Architecture

Most often in DNN-based image classification, raw images have been directly fed to the model, and the model extracts the global features and classifies the images based on the global features. However, instead of directly utilizing the raw images as input, the extracted frequency-domain information can be used as the input of the DNN model for image classification. To do so we have created our overall classifier model as described in Figure 6.2.

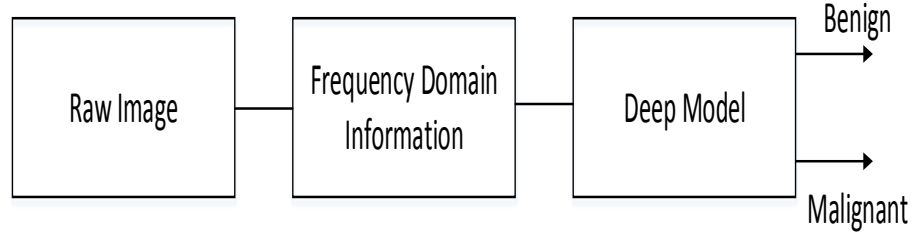


Figure 6.2: Overall Architecture for Breast-image classification

For the frequency-domain information extraction, we have selected one of

- Discrete Fourier Transform (DFT)
- Discrete Cosine Transform (DCT)

DFT for feature selection

The Fourier transform represents the signal in the frequency domain and calculates the most important information from the signal. Let $f(x, y)$ be the two-dimensional discrete signal, in this case the image. The Fourier transform of this signal can be represented as

$$\mathcal{F}(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \exp \frac{-(j2\pi ux)}{M} \exp \frac{(-j2\pi vy)}{N} \quad (6.1)$$

where $u = 0, \dots, M-1$ and $v = 0, \dots, N-1$, The original images can be recovered from the previous equation by applying the inverse Fourier transform:

$$f(x, y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \mathcal{F}(u, v) \exp \frac{(j2\pi ux)}{M} \exp \frac{(j2\pi vy)}{N} \quad (6.2)$$

where $x = 0, \dots, M-1$ and $y = 0, \dots, N-1$. Finding the DFT is computationally complex and time-consuming. The fast Fourier Transform (FFT) is an algorithm which solves the DFT with less computational complexity; we have utilised the FFT algorithm to extract the Fourier-transform coefficients.

The FFT values have been calculated from each image channel, then we selected the top "h" FFT coefficients and concatenated the features together. Now the total feature set

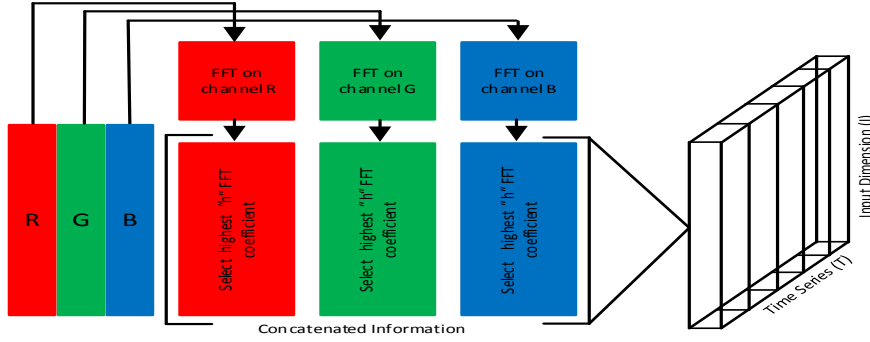


Figure 6.3: Feature preparation for LSTM/GRU model when FFT is utilised

will be size $F_{DFT} = 3 \times h$ and is a single-row vector. This single-row vector has been converted to a two-dimensional feature matrix $F_{FFT}^{TI} = \text{Time Series (T)} \times \text{Input Dimension (I)}$.

DCT for feature selection

The DCT method was first introduced in 1974 [228], and works for both 1D and 2D signals. DCT-II is a commonly used method for image processing and is also known as even-symmetric DCT. For our analysis, we have used DCT-II methods. For an image $f(x, y)$ the DCT can be represented as

$$\mathcal{D}(u, v) = \frac{2l(u)l(v)}{\sqrt{(MN)}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cos \left[\frac{(2x+1)u\pi}{2M} \right] \cos \left[\frac{(2y+1)v\pi}{2N} \right] \quad (6.3)$$

where $u = 0, \dots, M-1$ and $v = 0, \dots, N-1$, and

$$l(k) = \begin{cases} \frac{1}{2} & \text{if } k = 0 \\ 1 & \text{otherwise} \end{cases}$$

Now the original signal can be regained by the inverse DCT transform using

$$f(x, y) = \frac{2}{\sqrt{(MN)}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} l(u)l(v) \mathcal{D}(u, v) \cos \left[\frac{(2x+1)u\pi}{2M} \right] \cos \left[\frac{(2y+1)v\pi}{2N} \right] \quad (6.4)$$

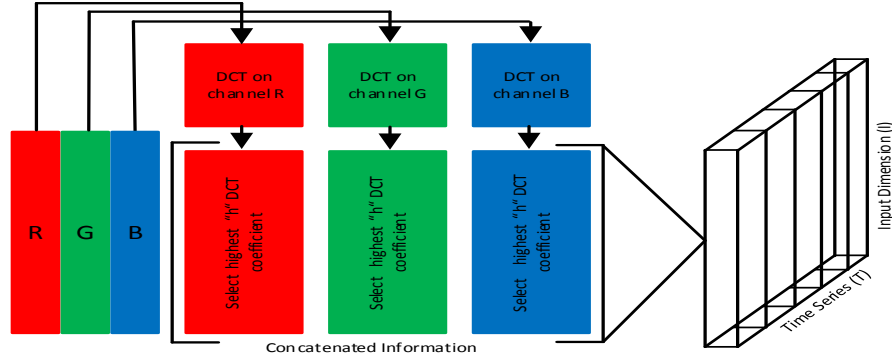


Figure 6.4: Feature preparation for LSTM when DCT is utilised

DCT values have been calculated from each image channel and the top "h" DCT coefficients selected and then the features concatenated together. Now the total feature set will be $F_{DCT} = 3 \times h$ and is a single-row vector. This single-row vector has been converted to a two-dimensional feature matrix $F_{DCT}^{TI} = \text{Time Series (T)} \times \text{Input Dimension (I)}$.

6.4 LSTM and GRU Methods

Learning from scratch might not be as fruitful as learning from a reference. This perception is realistic for human thinking, as the human brain always recalls references from previous learning. In the traditional NN the model always learns from scratch. However, a Recurrent Neural Network (RNN) feeds back the output information to the input. Figure 6.5 shows two diagrams which illustrate the differences between the conventional NN and an RNN. Let $\mathbf{X} = \{x_i\}$, $\mathbf{H} = \{h_j\}$ and $\mathbf{Y} = \{y_k\}$ represent the input, hidden and output layers where $i \in \{1, 2, 3, \dots, Q\}$, $j \in \{1, 2, 3, \dots, R\}$ and $k \in \{1, 2, 3, \dots, P\}$. The output layer of an RNN model can be expressed as

$$y_t = W_{ho}h_t$$

where h_t is defined as

$$h_t = \sigma(W_{hh}h_{t-1} + W_{ih}x_t + b_h) \quad (6.5)$$

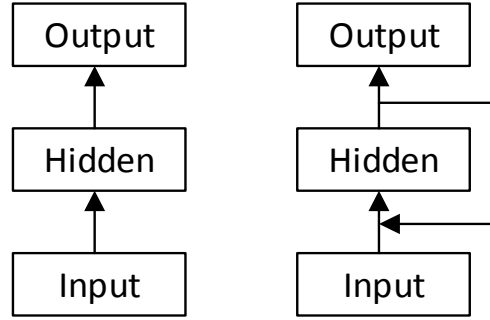


Figure 6.5: Conventional NN and RNN models

Here

$W_{i,h}$ represents the weight matrix from the input to the hidden layer;

$W_{h,h}$ represents the weight matrix from the hidden to the hidden layer;

$W_{h,o}$ represents the weight matrix from the hidden to the output layer;

Conventional RNN suffers due to long term-memory dependencies. To overcome this problem, Hochreiter et al. proposed the Long Short-Term Memory (LSTM) architecture which is an advanced version of the RNN model [209]. The LSTM layer contains a forget gate which controls the flow of information. Figure 6.6 represents a cell structure of a LSTM network. The main parameters of the LSTM network can be represented as:

$$i_t = \tanh(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (6.6)$$

$$j_t = \sigma(W_{xj}x_t + W_{hj}h_{t-1} + b_j) \quad (6.7)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (6.8)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (6.9)$$

$$c_t = c_{t-1} \odot f_t + i_t \odot j_t \quad (6.10)$$

$$h_t = \tanh(c_t) \odot o_t \quad (6.11)$$

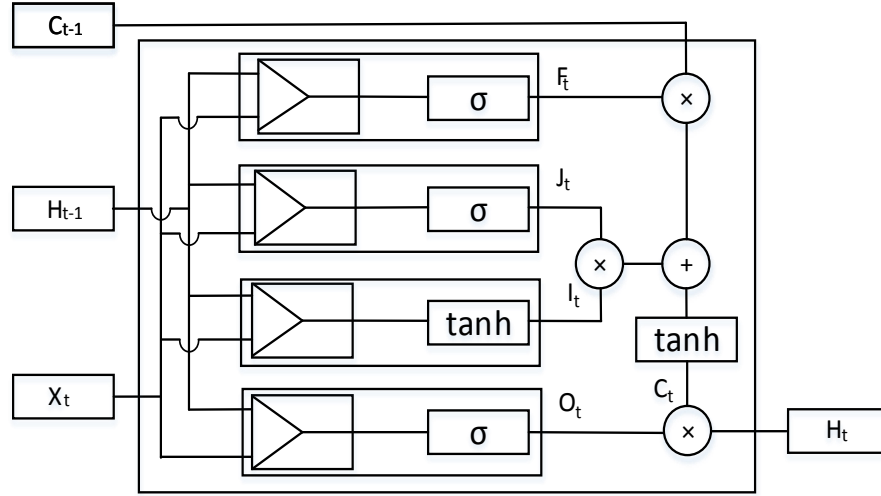


Figure 6.6: A generalised cell structure of an LSTM

where f_t is the forget gate, i_t is the input gate, h_t provides the output information and c_t represents the cell state [210]. Here all $W_{\times\times}$ and b_{\times} represent the corresponding layer's weight matrix and bias vectors. Another advancement of the LSTM network is the Gated Recurrent Unit (GRU). In the GRU model the forget gate and the input gate are merged, and the Hidden Gate and the Cell state also merge, with some other modifications.

6.4.1 Our Model

We have utilised the LSTM model as well as the GRU model separately for the data classification task. Figure 6.7 shows the model which has been used for the classification. The hidden vector and the cell state go to the next adjacent layer. However, like a normal NN we also stacked a number of hidden cell (LSTM/GRU) layers one after another, so that each cell (LSTM/GRU) of the first layer produces a hidden vector with cardinality 82, and layer-2, layer-3 and layer-4 produce hidden vectors with cardinality 42. The output of the last cell of layer-4 (upper top-right corner) passes through a dropout layer, which drops out 20 percent of the neuron information. Then two consecutive dense layers

are stacked one after another.

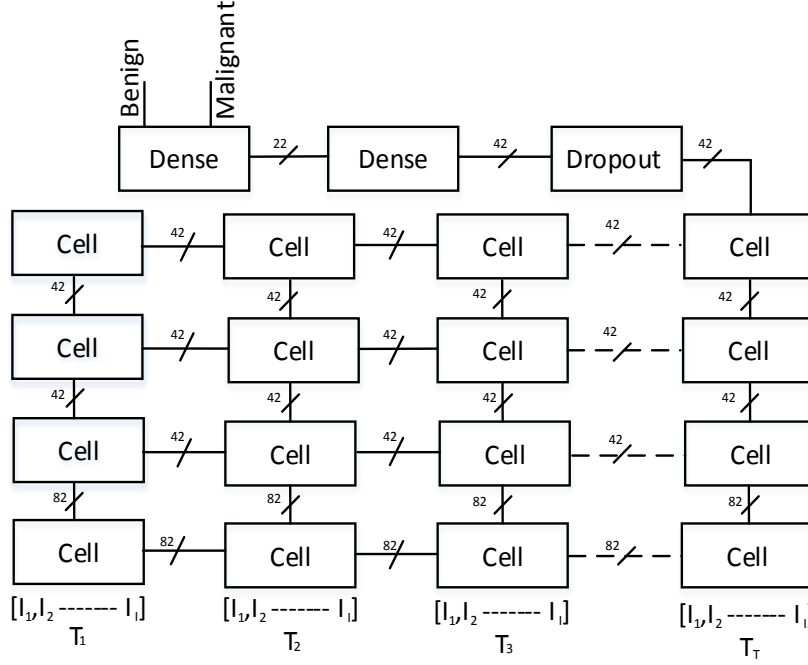


Figure 6.7: The proposed model where the cells (LSTM/GRU) are stacked

Both LSTM and GRU take the input which contains the Time Series (T) and Input Dimension (I). Based on the values of T, I and the method LSTM and GRU, we have divided our experiment into eight different cases, summarized in Table 6.1.

6.5 Results and Discussion

We have utilised the BreakHis breast-image dataset which contains four sets of images "m×" where $m = \{40, 100, 200, 400\}$ and × represents the magnification factor. For Case-1, Case-2, Case-3 and Case-4, the value of Time Series (T) and Input dimension (I) is fixed at 31 and 93; for the rest of the cases the values of T and I are fixed at 93 and 31 respectively.

The first quarter of Table 6.2 shows the performance for the different Cases, where the

Table 6.1: Description of the Cases

Case	Method	Features	T×I
Case-1	LSTM	FFT	31×93
Case-2	LSTM	DCT	31×93
Case-3	GRU	FFT	31×93
Case-4	GRU	DCT	31×93
Case-5	LSTM	FFT	93×31
Case-6	LSTM	DCT	93×31
Case-7	GRU	FFT	93×31
Case-8	GRU	DCT	93× 31

40× dataset has been utilised. For Case-3, where the GRU model has been used along with FFT features, 84.97% Accuracy has been achieved. In this particular case the Recall value is 92.82%; this indicates that 9.17% of the Malignant images have been mis-classified to Benign images. On the other hand the Specificity value is 70.68% and the False Positive Rate (FPR) is 29.31%. This indicates that this particular case is moderately sensitive and less specific.

Information of the second quarter of Table 6.2 shows the performance of the classifier models on the 100× dataset. In this scenario, when we use Case-8 the achieved accuracy is 87.80%. In this particular case the Recall value is 92.27%; this indicates that 7.70% of the Malignant data has been mis-classified as Benign data. On the other hand the Specificity value is 76.51% and the False Positive Rate (FPR) is 23.42%. This indicates that this particular situation is less specific. The best Specificity 80.00% has been achieved on the 100× dataset utilizing Case-7. In this particular situation the accuracy is 86.17%.

Information in the third quarter of Table 6.2 shows the performance when we utilised

Table 6.2: Performances of the different Cases on different datasets

	Case	Accuracy (%)	Specificity (%)	FPR (%)	FNR (%)	Recall (%)	Precision (%)	F-Measure (%)
40×	Case-1	84.47	76.00	24.00	12.00	88.00	90.00	89.00
	Case-2	81.47	74.13	25.86	15.52	84.47	89.00	81.00
	Case-3	84.97	70.68	29.31	09.17	90.82	88.00	90.00
	Case-4	74.26	82.18	17.80	28.47	71.52	91.00	80.00
	Case-5	70.95	00.00	100.00	00.00	100.00	71.00	83.00
	Case-6	84.47	66.01	33.90	08.00	99.20	87.00	89.00
	Case-7	84.47	74.13	25.86	11.26	88.70	89.00	89.00
	Case-8	83.80	73.56	26.43	12.00	88.00	89.00	89.00
100×	Case-1	85.69	74.85	25.14	10.00	90.00	90.00	90.00
	Case-2	87.32	78.85	21.14	09.31	90.68	92.00	91.00
	Case-3	85.36	78.28	21.71	11.81	88.18	90.00	91.00
	Case-4	83.25	67.42	32.57	10.45	89.44	87.00	88.00
	Case-5	71.50	0.00	100.00	0.00	100.00	72.00	83.00
	Case-6	71.54	0.00	100.00	0.00	100.00	72.00	83.00
	Case-7	86.17	80.00	20.00	11.36	88.63	92.00	90.00
	Case-8	87.80	76.51	23.42	7.70	92.27	91.00	92.00
200×	Case-1	85.69	84.34	15.65	04.60	95.33	93.00	94.00
	Case-2	91.74	91.91	08.08	08.35	91.64	96.00	94.00
	Case-3	92.90	84.34	15.65	04.42	95.57	93.00	94.00
	Case-4	89.09	84.34	15.65	06.14	93.85	90.00	92.00
	Case-5	67.27	0.00	100.00	00.00	100.00	67.00	80.00
	Case-6	67.27	0.00	100.00	00.00	100.00	67.00	80.00
	Case-7	90.74	81.31	18.68	4.66	95.33	91.00	93.00
	Case-8	93.05	87.87	12.12	4.42	95.57	94.00	95.00
400×	Case-1	83.88	85.24	14.75	16.80	83.19	92.00	87.00
	Case-2	88.46	83.60	16.93	08.81	91.18	91.00	91.00
	Case-3	85.71	74.86	25.13	08.81	91.18	88.00	89.00
	Case-4	83.25	84.15	15.84	07.71	92.18	92.00	92.00
	Case-5	66.48	00.00	100.00	00.00	100.00	66.00	80.00
	Case-6	66.48	00.00	100.00	00.00	100.00	66.00	80.00
	Case-7	88.46	81.42	18.57	07.98	98.02	91.00	91.00
	Case-8	89.19	88.52	11.47	10.46	89.53	94.00	92.00

the $200\times$ dataset. On the dataset $200\times$ when we used Case-8, the achieved accuracy is 95.57% and the False Negative Rate (FNR) is 4.42%; this indicates that this particular scenario is highly sensitive. In this case 87.87% of the Benign images have been classified as Benign and 12.12% of Benign images misclassified as Malignant. In terms of F-measure this algorithm provides 95.00%. Overall this particular case shows moderate Specific performance and quite impressive Sensitive performance. Information in the last quarter of Table 6.2 shows the performance of the different cases on the $400\times$ dataset. For Case-8, the achieved accuracy is 89.19%. The recall and the specificity almost maintain the same values, at around 89.00%. This indicates that around 11.00% of the Malignant data has been misclassified as Benign data and the reverse is also true. Case-5 utilises the LSTM method and FFT features, and the T and I values are fixed at 93 and 31. For all the datasets, Case-5 shows the FPR and Recall both are 100.00%; this indicates that all the data has been classified as Malignant data irrespective of their original class. For case-6, where LSTM and DCT features have been used and the I and T values are fixed at 91 and 31, when we apply this case on the $100\times$, $200\times$, and $400\times$ datasets all the data has been classified as Malignant irrespective of their original classes. However when we apply Case-6 on the $40\times$ data set the Accuracy, Specificity and Recall are 84.47%, 66.01% and 99.20% respectively. Overall Case-5 shows the worst performance irrespective of the dataset, and the performance of Case-6 is very poor with some exceptions. Figure 6.5 shows the number of parameters needed for the overall classification operation and the required time to run per epoch; Case-1 and Case-2 required the same time and parameters. Two things are common in these two cases: both utilised the LSTM method and the values of T and I are fixed at 31 and 93 respectively. However Case-1 utilises the FFT as feature and Case-2 utilises the DCT as feature. If we consider Case-5 and Case-6, they require the same parameters as well as the same time to perform in each epoch. Two things are common in these two cases: both utilised the LSTM method and the values of T and I are

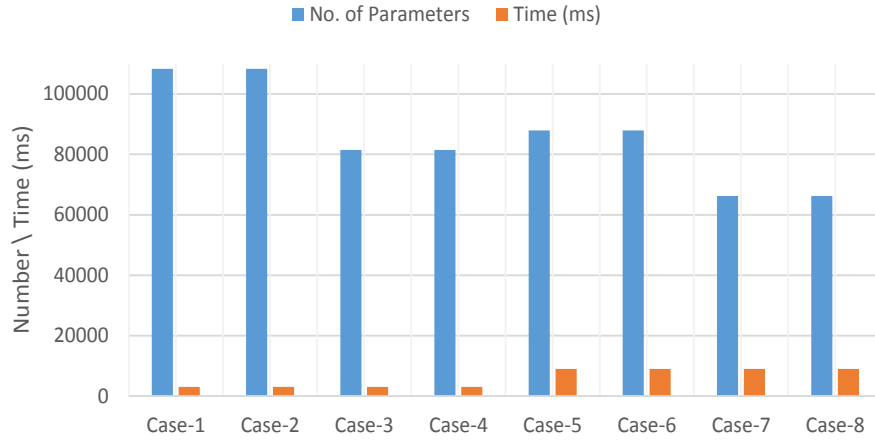


Figure 6.8: Number of parameters and required time for individual cases

fixed at 93 and 31 respectively. However Case-5 utilises the FFT as feature and Case-6 utilises DCT as feature. The number of parameters required for Case-5 and Case-6 is less than the number of parameters required to perform Case-1 and Case-2, however both Cases 4 and 5 require more time to perform per epoch than Case 1 and Case 2. One thing common to Cases 1, 2, 5 and 6 is that all utilised the LSTM method, however when they utilise T equal to 31 they require more parameters but less time, and when they utilise T equal to 93 they require more time but fewer parameters. Case-3 and Case-4 require the same parameters to perform the operation and the same time for each epoch. Two things are common in these two cases: both utilised the GRU method and T and I are fixed at 31 and 93 respectively. However Case-1 utilises FFT as feature and Case-2 utilises DCT as feature. If we consider Case-7 and Case-8, they require same parameters as well as the same time to perform at each epoch. Two things are common in these two cases: both utilised the GRU method and T and I are fixed at 93 and 31 respectively. However Case-7 utilises FFT as feature and Case-8 utilises the DCT as feature. The GRU method has been utilised by Cases 3, 4, 7 and 8. however when they utilise T equal to 31 they require more parameters but less time and but when they utilise T equal to 93 they require more time fewer parameters. For Case-5, irrespective of the dataset it

provides the worst performances among all the other cases. Specially Case-5 classifies all the data as Malignant irrespective of their original class. Figure 9 shows the Accuracy, Loss, Kullback–Leibler Divergence (KLD) and Matthews Correlation Coefficient (MCC) performance for Case-5, which was performed on the $200\times$ dataset.

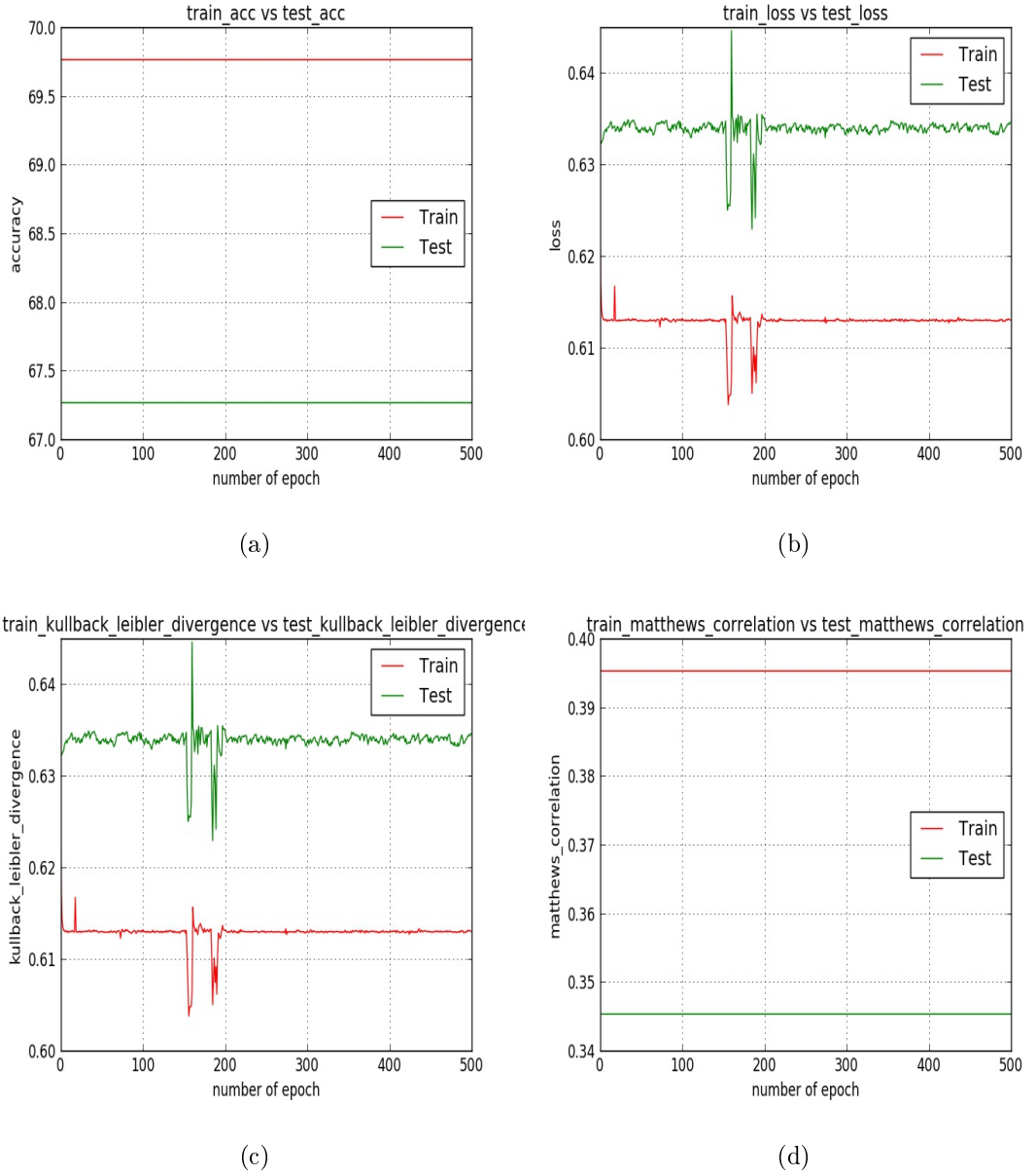
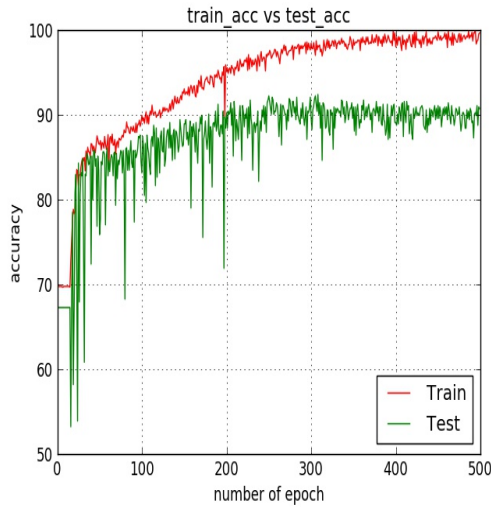
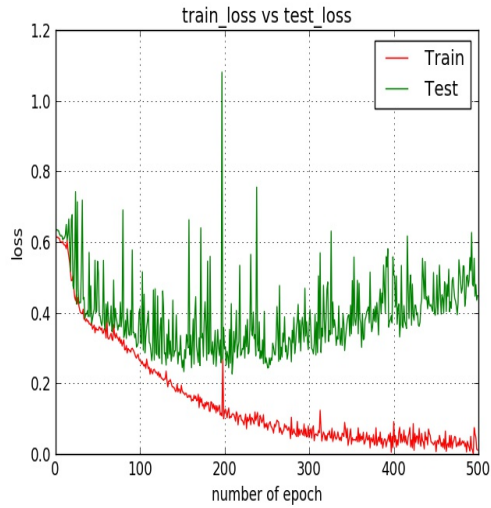


Figure 6.9: (a), (b), (c) and (d) show the Accuracy, Loss, Kulback Divergence and Matthews correlation coefficient values for Case-5 on the $200\times$ dataset

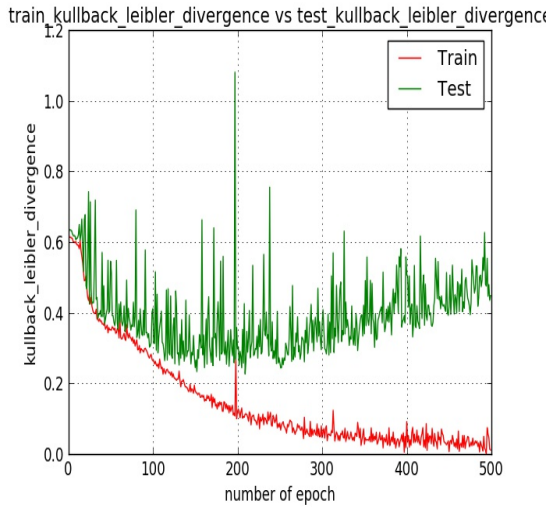
Among all the available cases, Case-7 and Case-8 require fewer parameters, however Case-1 and Case-2 require less time. This indicates that if we use the LSTM method along with the lower value of T this requires less time, however if we utilise the GRU value along with the higher value of T it requires fewer parameters but more time.



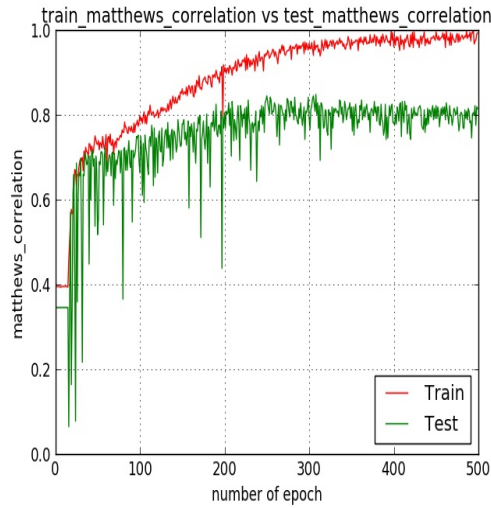
(a)



(b)



(c)



(d)

Figure 6.10: (a), (b), (c) and (d) show the Accuracy, Loss, Kulback Divergence and Matthews correlation coefficient values for Case-8 on the $200\times$ dataset

In general, among all the cases, Case-8 gives better performance on all the datasets. In particular, the best performance is achieved when we utilised Case-8 and the $200\times$ dataset. Figure 10 represents the Accuracy, loss, KLD and MCC performance over epochs 0 to 500. Up to around epoch 50 both the train and test accuracy remain almost the same. After 50 epochs the train accuracy outperforms the test accuracy. After around epoch 250 the test accuracy remains a constant, but as the epoch goes higher, the train accuracy still shows better performance. After some initial epochs, the difference between the train and test loss increases and as the epoch goes on this loss difference continuously increases. Figure c shows the KLD and figure d represents the MCC. As we know, the MCC varies from -1 to +1. For the training case the MCC values are almost the highest values, however the test MCC values were around 0.8.

6.6 Conclusion

In this chapter, we have classified the BreakHis breast image dataset based on a DNN model, created using a stack of the LSTM and GRU models. The best result has been achieved when we have utilised the GRU model. the best performance is achieved when we use Time Series T, Input Dimension Value I equal to 93 and 31 respectively. In this particular case the achieved Accuracy, Recall, Precision values are 93.05%, 95.57% and 94.00% respectively. This is the best available performance on this dataset.

Chapter 7

Histopathological Breast-Image Classification using Local and Frequency domains by Convolutional Neural Network

7.1 Abstract

Identification of the malignancy of tissues from Histopathological images has always been an issue of concern to doctors and radiologists. This task is time-consuming, tedious and moreover very challenging. Success in finding malignancy from Histopathological images primarily depends on long experience, though sometimes experts disagree on their decisions. However, Computer Aided Diagnosis (CAD) techniques help the radiologist to take

Published as: A. A. Nahid, Y. Kong, “Histopathological Breast-Image Classification Using Local and Frequency Domains by Convolutional Neural Network”, *Information, MDPI*, vol. 9, no. 1, pp. 1–26, 2018. [Online].

a second opinion which can increase the reliability of the radiologist,s decision. Among the different image analysis techniques, classification of the images has always been a challenging task. Due to the intense complexity of biomedical images, it is always very challenging to provide a reliable decision about an image. The state-of-the-art Convolutional Neural Network (CNN) technique has had great success in natural image classification. Utilizing advanced engineering techniques along with the CNN, in this chapter we have classified a set of Histopathological Breast-Cancer (BC) images utilizing a state-of-the-art CNN model containing a residual block. Conventional CNN operation takes raw images as input and extracts the global features, however the object oriented local features also contain significant information, for example the Local Binary Pattern (LBP) represents the effective textural information, Histogram represent the pixel strength distribution, Contourlet Transform (CT) gives much detailed information about the smoothness about the edges, Discrete Fourier Transform (DFT) derives frequency-domain information from the image. Utilizing these advantages, along with our proposed novel CNN model, we have examined the performance of the novel CNN model as Histopathological image classifier. To do so we have introduced five cases: a) Convolutional Neural Network Raw Image (CNN-I), b) Convolutional Neural Network CT Histogram (CNN-CH), c) Convolutional Neural Network CT LBP (CNN-CL), d) Convolutional Neural Network Discrete Fourier Transform (CNN-DF), e) Convolutional Neural Network Discrete Cosine Transform (CNN-DC). We have performed our experiments on the BreakHis image dataset. The best performance is achieved when we utilize the CNN-CH model on a $200\times$ dataset which provides Accuracy, Sensitivity, False Positive Rate, False Negative Rate, Recall Value, Precision and F-measure of 92.19%, 94.94%, 5.07%, 1.70%, 98.20%, 98.00 and 98.00 % respectively.

7.2 Introduction

Cancer, being a serious threat to human life, is actually a combination of diseases, and more specifically unwanted and abnormal growth of the cells of the human body is known as cancer. Cancer can attack any part of the body and can then be distributed to any other part. Different types of cancer exist, but among all the cancers women are more vulnerable to Breast Cancer (BC) than men, because of the anatomical structure of women. Statistics show that each year more people are newly affected by BC, at an alarming rate. Figure 7.1 shows the number of females newly facing BC as well as the number of females dying since the year 2007 in Australia. This figure shows that more and more females are newly facing BC, and the number of females dying of it has also increased in each year. This is the situation of Australia (population 20 - 25 million), but it can be used as a symbol of the BC situation of the whole world.

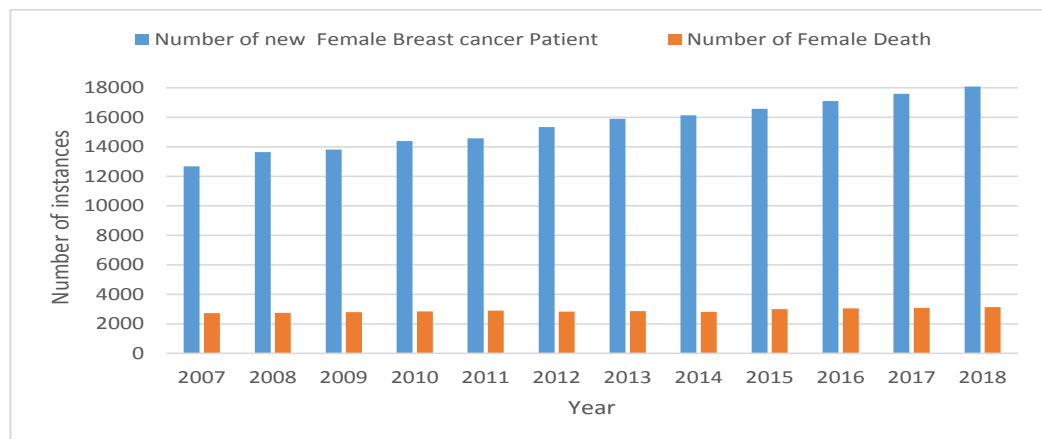


Figure 7.1: New cases of breast cancer for women and number of women dying in the last twelve years

Proper investigation is the first step in proper treatment of any disease. Investigation of BC largely depends on investigation of biomedical images such as Mammogram, MRI, Histopathological etc. Manual investigation of this kind of images largely depends on the

expertise of the doctors and physicians. As humans are error prone, so even an expert can give wrong information about the diagnostic images. Beside this, biomedical image investigation always requires a large amount of time. However CAD techniques are largely utilized for biomedical image analysis such as cancer identification and classification. The use of CAD allows the patient and doctor to take a second opinion.

Different biomedical image-analysis techniques are available and different research groups have investigated the identification and classification of BC. The conventional image-classification techniques such as Support Vector Machines (SVM), Random Forest (RF), Bayesian classifier etc. algorithms have been largely utilized for the image classification. Utilizing an SVM, a set of cancer images was first classified by A. Bazzani et al. and their findings have been compared with the Multi Layer Perception (MLP) technique [229]. I. Naqa et al. [124] utilized the kernel method along with SVM techniques for better performance for the classification, where they obtained around 93.20% accuracy. A set of Histopathological images has been classified using Scale Invariant Feature Transform (SIFT) and Discrete Cosine Transform (DCT) features with an SVM for classification by N. Mhala et al. [130]. Law's Texture features have been utilized for Mammogram (322 images) image classification and 86.10% accuracy obtained by J. Dheeba. et al. [140]. M. Taheri et al. [141] utilized intensity information, Auto Correlation Matrix and Energy values for breast-image classification and obtained 96.80% precision and 92.50% recall with 600 Mammogram images. A set of ultrasound images have been classified by F. Shirazi et al. [138], where Regions of Interest (ROI) have been extracted for reduction of the computational complexity. J. Levman et al. [135] classify a set of MRI images (76 images) into benign and malignant classes, utilizing Relative Signal Intensities and Derivative of Signal Strength as features.

The RF method has also been used for image classification. A set of Mammogram images has been classified by S. Angayarkanni et al. [118] and they achieved 99.50% accuracy

using the Gray-Level-Cooccurrence Matrix (GLCM) as feature. G. Gatuha et al. [230] utilized Mammogram images for image classification using a total of 11 features and achieved 97.30% accuracy. Breast Histopathological images have been classified by Y. Zhang et al. [117] and they achieved 95.22% accuracy, where they utilized the Curvelet Transform, GLCM, and Completed Local Binary Pattern (CLBP) methods for feature extraction. GLCM and Gray-Level-Run-Length-Matrix (GLRLM) have been utilized along with the RF algorithm by J. Diz et al. [113] for Mammogram image classification with 76.60% accuracy. The Bayes method has also been used for image classification. E. Kendall et al. [151] utilized the Bayes method for Mammogram image classification with the DCT method for feature selection. Their obtained sensitivity was 100.00% and specificity was 64%. Statistical and Local Binary Pattern (LBP) features along with the Bayesian method have been utilized by F. Claridge et al. [153] on two Mammogram image sets. When they used the MIAS dataset their best achieved accuracy was 62.86%.

Other than RF, SVM, Bayes method, the Neural Network (NN) method have largely been utilized for image classification. K. T. Rajakeerthana et al. [50] classified a set of Mammogram images and obtained 99.20% accuracy. Thermographic images have been classified by V. Lessa et al. [51] and they utilized NN method along with a few statistical values such as mean, median, skewness, kurtosis, median as features and obtained 85.00% accuracy with a specificity value of 83.00%. Haralick and Tamura features have been utilized by W. Peng et al. [57] along with an NN network. They used Rough-Set theory for the feature reduction. S. Silva et al. [63] utilized 22 different morphological features such as convexity, lobulation index, elliptic normalized skeleton along with NN for ultrasound image classification and obtained 96.98% accuracy. E. Melendez et al. [65] utilized Area, Perimeter, Circularity, Solidity etc. along with NN and achieved sensitivity and specificity of 96.29% and 99.00%.

As the literature shows, different methods and techniques have been utilized for image

classification on different breast-image datasets using different image-classification techniques. However the state-of-the-art image classification technique of the Convolutional Neural Network (CNN) has put its strong footprint in the image-analysis field, specially the image-classification field. Though the model "AlexNet" proposed by A. Krizhevsky has gained a new momentum in the CNN research field, a CNN model was first utilized by K. Fukushima et al. [79]. who proposed the "Neocognitron" model which recognises stimulus patterns. For the mammogram image classification Y. Wu et al. first utilized the CNN model [80]. Though little work on the CNN model had been done to the end of the 20th century, this model has only gained momentum from the AlexNet model. Advanced engineering techniques have been used by research groups such as the Visual Geometry Group and Google, which have modeled the VGG-16, VGG-19 and GoogleNet models. J. Arevalo et al. [93] classified benign and malignant lesions using the CNN model, and this experiment was performed on 766 mammogram images, where 426 images contain benign and 310 malignant lesions. Before classifying the data they utilized preprocessing techniques to increase the image enhancement and obtained a 0.82 ± 0.03 ROC value. GoogleNet and AlexNet methods have been utilized by M. Zejmo et al. [98] for the classification of cytological specimens into benign and malignant classes. The best accuracy obtained when they utilized the GoogleNet model was 83.00%. Y. Qiu et al. [101] used the CNN method to extract global features for Mammogram image classification and obtained an average achieved accuracy of 71.40%. S. Fotin et al. also utilized the CNN method for tomosynthesis image classification and obtained an AUC curve value of 0.93. Transfer learning is another important concept of the CNN method which allows the model to not extract features from scratch, rather applying a weight-sharing concept to train a model. This method is helpful when the database contains fewer images. F. Jiang et al. [99] utilized a transfer learning method for Mammogram image classification and obtained an AUC of 0.88. Before utilizing it in a CNN model they performed a preprocessing opera-

tion to enhance the images. S. Suzuki et al. [100] also used the benefit of transfer learning techniques to train their model to classify mammogram images and obtained sensitivity 89.9%. They performed their experiment with only 198 images.

Most image classification based on the CNN method has been performed based on global feature-extraction techniques. Recently researchers have also shown an interest in how local features can be utilized with the CNN model for data classification. Both global and local features have been utilized by H. Rezaeilouyeh et al. [226] for Histopathological image classification. For local feature extraction the authors utilized the Shearlet transform and obtained an accuracy of $86 \pm 3.00\%$. For local feature extraction K. Sharma et al. [95] used the GLCM, GLDM methods and then fed the local features to a CNN model for the Mammogram image classification, obtaining 75.33% accuracy for the fatty and dense tissue classification. Both global and local features have been used by Z. Jiao et al. [97] for mammogram image classification and they obtained 96.70% accuracy. T. Kooi et al. [103] utilized both global features and hand-crafted features for Mammogram image classification. In their experiment they also utilized the transfer learning method.

The Contourlet Transform (CT) has been used for image analysis. Using CT, the distribution of Mammograms (MIAS dataset) has been calculated by S. Anand et al. [231]. Along with GLCM and morphological features, CT features have been utilized for the Mammogram image classification with the SVM method, and obtained a mean Accuracy around 100.00% by F. Moayedi et al. [232]. The non-subsampled CT transform has been utilized for Breast mass classification by J. S. Leena Jasmine along with the SVMs techniques [233]. Fatemeh Pak et al. also utilized Non-subsampled CT for breast-image (MIAS dataset) classification and obtained 91.43% mean Accuracy and 6.42% mean FPR [234].

Inspired by the usefulness of local-features utilization techniques with the CNN, this chapter has also classified a set of Histopathological images (BreakHis dataset) using lo-

cal features along with the CNN model. For the local-feature selection we have utilized the CT transform, LBP and Histogram information. We have also extracted frequency-domain information and tried to find how the CNN model behaves when we provide frequency-domain information. To do so we have organized our chapter as follows, Section 7.2 describes related research, Section 7.3 describes the overall architecture for the image classification, Section 7.4 describes the feature-extraction and data-preparation techniques, Section 7.5 describes the novel Convolutional-Neural-Network (CNN) model, Section 7.6 describes the performance measuring parameters, Section 7.7 describes the performance of our model on the BreakHis dataset as well as compare with the present findings, and we conclude our chapter in Section 7.8.

7.3 Overall Architecture

Benign and Malignant image classification has always been a challenging task. The level of complication of the data classification increases when we consider Histopathological images, as an example the left side Figure 7.2 represents the Benign and the right side figure represent the Malignant images. Every supervised classification technique follows a predefined working mechanism, such as selection of dataset, features and model to perform the classification, then a set of performance measuring parameters is tested based on model performance parameters. The selected dataset is normally split into train and test datasets. A hypothetical model is established based on the training dataset, and later this hypothetical model's performance is evaluated by the test dataset.

Conventionally, handcrafted features or local features are extracted and utilized for the input of a classifier model. However, in most of the work using CNN-based image classification, raw images are fed directly to the CNN model. From the raw images, the CNN model tries to extract features globally. In this work we have utilized raw images as

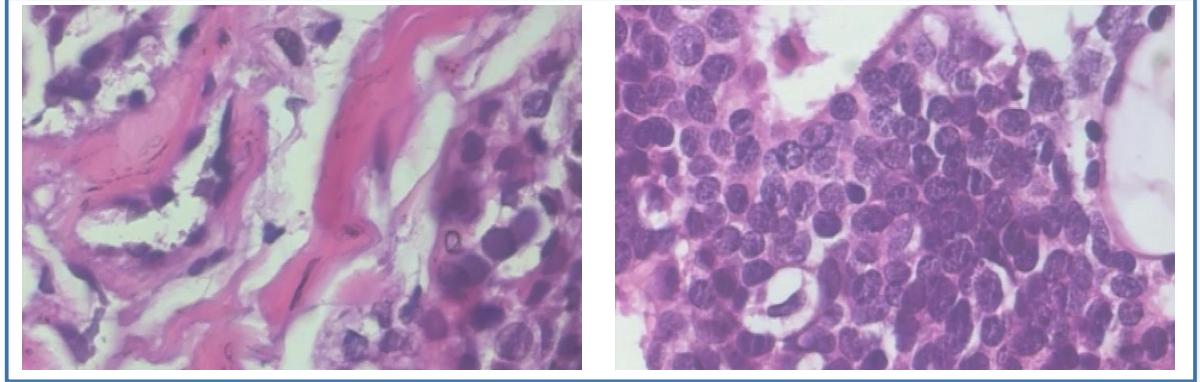


Figure 7.2: Left side represents Benign and right side Malignant histopathological images (This data has been collected from the BreakHis dataset)

well as descriptive handcrafted local features and frequency-domain information for the image classification along with the CNN model. Figure 7.3 shows the overall classifier model which has been used for the data classification.

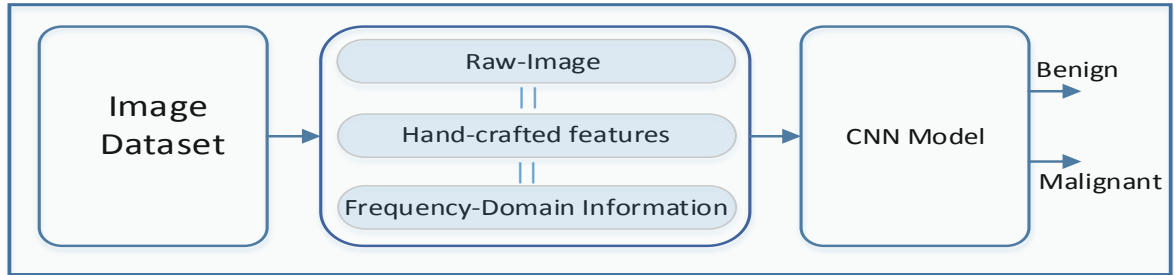


Figure 7.3: Overall image-classification model.

Based on how we prepare the features to feed them in to the CNN model, we have divided our work into the following three cases:

- Case1: In this case, the image has been directly fed to the CNN model which is named CNN-I. To reduce the complexity we have reshaped each of the original images of the dataset to a new image matrix of size $\mathcal{H}_I = h_1 \times h_2 \times C$, where C represents the number of channels. As we have utilized RGB images, the value of

$C = 3$.

- Case2: Case2 utilizes local descriptive features which have been collected through the Contourlet Transform (CT), Histogram information and Local Binary Pattern (LBP). Case2 is further divided into two sub cases:
 1. Case2a: Selected statistical information has been collected from the CT coefficient data and this statistical data has been further concatenated with the Histogram information. This case has been named CNN-CH. The feature matrix for each of the images is represented as $\mathcal{H}_{CH} = h_1 \times h_2$.
 2. Case2b: Selected statistical information has been collected from the CT coefficient data and this statistical data has been further concatenated with the LBP. This case has been named CNN-CL. The feature matrix for each of the images is represented as $\mathcal{H}_{CL} = h_1 \times h_2$.
- Case3: Case3 utilizes frequency-domain information for the image classification, collected using the Discrete Fourier Transform (DFT) and the Discrete Cosine Transform (DCT). This case has been further subdivided into two sub-cases:
 1. Case3a: DFT coefficients have been utilized as an input for the classifier model, named CNN-DF. The feature matrix for each of the images is represented as $\mathcal{H}_{DF} = h_1 \times h_2$.
 2. Case3b: DCT coefficients have been utilized as an input for the classifier model, named CNN-DC. The feature matrix for each of the images is represented as $\mathcal{H}_{DC} = h_1 \times h_2$.

7.4 Feature Extraction and Data Preparation

We have utilized three cases to analyse our data. Case1 or CNN-I directly feeds the raw data to the CNN model for further analysis. However Case2 and Case3 utilize handcrafted features with CT, Histogram, LBP, Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT).

7.4.1 Data Preparation for Case2

For Case2, we have extracted a set of statistical information utilizing the value of CT coefficients which has been collected after applying CT to each of the images. A CT is an extension of the Wavelet Transform (WT). WT ignores the smoothness along the contour and it provides less directional information about the images, whereas CT overcomes this problem of WT and gives better information about the contour and direction edges of an image [235]. CT method utilizes multi-scale Laplacian Pyramid (LP) and a Directional Filter Bank (DFB).

- Laplacian Pyramid (LP): The image pyramid is an image-representation technique where the represented image contains only relatively important information. This technique also produces a series of replications of the original images, but those replicated images have less resolution. A few pyramid methods are available such as Gaussian, Laplacian and WT. Burt and Adelson introduced the Laplacian Pyramid (LP) method. In the case of CT, the LP filter decomposes the input signal into a coarse image and a detailed image (bandpass image) [236]. Each bandpass is further processed and the bandpass directional sub-band signals calculated.
- Directional Filter Bank: A DFB sub-divides the input image into 2^{n+1} sub-bands. Each of the sub-bands has a wedge-shaped frequency response. Figure 7.4 a shows the wedge-shaped frequency response for a 4-band response.

Let the input image $\mathcal{J}(x, y)$ feed to the LP filter (LP_n), where $n = 1, 2, \dots, N$, which decomposes $\mathcal{J}(x, y)$ images into the low-pass signal $\mathcal{L}_n(x, y)$ and the detailed signal $\mathcal{T}_n(x, y)$. The detailed image $\mathcal{T}_n(x, y)$ is passed through a DFB to get the directional images. In general the detailed image at level j , $\mathcal{T}_j(x, y)$ is further decomposed by DFB into 2^j -level directional images $\mathcal{C}_{j,k}(l, j)$. Figure 7.4 b shows an overall CT procedure.

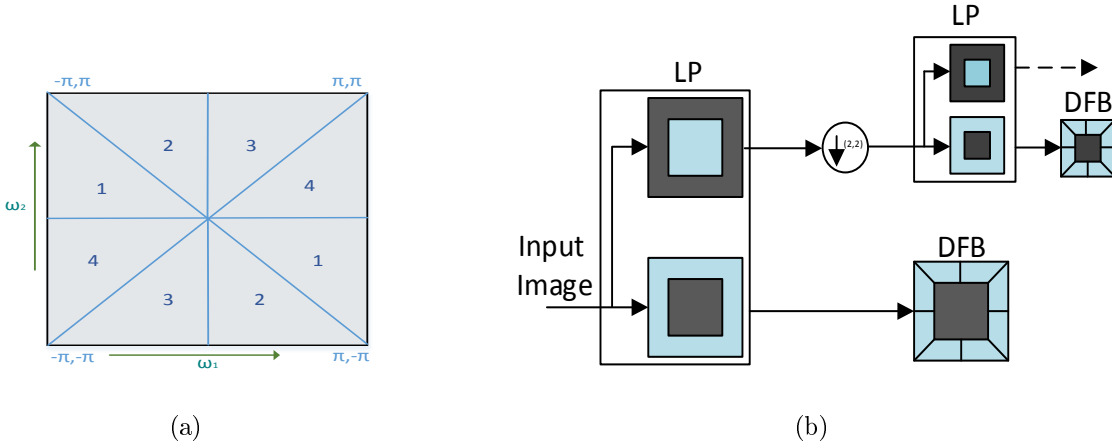


Figure 7.4: (a) Wedge-shaped frequency response for 4-band decomposition and (b) Contourlet Transform working mechanism

As CT is an iterative operation, it continuously produces low-pass signals and directional signals into some predefined level. Among the available lowpass signals and the directional signals we have deliberately selected a set (cardinality of the set is sixteen) of statistical features:

- Maximum Value (MA)
- Minimum Value (MI)
- Kurtosis (KU)
- Standard Deviation (ST).

The CT operation has been performed on each of the image channels individually. For a single channel of each of the input images, we calculated sixteen MA, sixteen MI, sixteen KU, sixteen ST values which have used as features. Features extracted from the single channel utilizing the CT and statistical method can be written as $\mathcal{F}_{CS} =$

$\{16 \times \text{MA} + 16 \times \text{MI} + 16 \times \text{KU} + 16 \times \text{ST}\}$, so a single channel image produces sixty-four feature values using CT and statistical information. As our images are RGB, we have utilized Red, Green and Blue channels, so the total number of features due to the CT utilization will be $\mathcal{F}_{CT} = 3 \times \mathcal{F}_{CS}$.

Histogram Information

A graphical display which represents the frequency of each of the particular intensities in an image is known as a histogram. Let the feature set collected for the histogram information from a single channel be represented as \mathcal{F}_{HIS} . A single RGB image provides a total $\mathcal{F}_{HIT} = 3 \times \mathcal{F}_{HIS}$ features, where the cardinality of \mathcal{F}_{HIT} will be 768. As Case2a that is CNN-CH utilizes statistical information collected from CT as well as histogram information, the total concatenated features will be $\mathcal{F}_{C2a} = \{\mathcal{F}_{CT}, \mathcal{F}_{HIT}\}$, and cardinality of \mathcal{F}_{C2a} will be 960. We have added zero padding at the end of the feature set \mathcal{F}_{C2a} to reshape the \mathcal{F}_{C2a} vector to a 31×31 matrix, to produce the matrix \mathcal{H}_{CH} .

Local Binary Pattern

The Local Binary Pattern (LBP) is proposed by Ojala et al. [25] which represents an image $\mathcal{J}(x, y)$ by a two-dimensional matrix, where each entry of this newly created two-dimensional matrix is labeled by an integer. Basically this matrix represents a local pattern and structural distribution of the image information. A single channel provides 256 LBP features. Let the feature set collected for the LBP information from a single channel be represented as \mathcal{F}_{LBS} . A single RGB image provides a total $\mathcal{F}_{LBT} = 3 \times \mathcal{F}_{LBS}$ features, so the cardinality of \mathcal{F}_{LBT} will be 768. As Case2b that is CNN-CL utilizes statistical features from CT and LBP, so the total concatenated features will be $\mathcal{F}_{C2b} = \{\mathcal{F}_{CT}, \mathcal{F}_{LBT}\}$, with a cardinality of 960. We have added zero padding at the end of the feature set \mathcal{F}_{C2b} to reshape the \mathcal{F}_{C2b} vector to a 31×31 matrix, to produce the matrix

\mathcal{H}_{CL} .

7.4.2 Data Preparation for Case3

For Case3, we have utilized frequency-domain information as the features. To find the frequency-domain information we have utilized the DFT and DCT transforms.

DFT for feature selection

Frequency-domain information reveals valuable information from the signal which can be extracted using the Fourier Transform. This frequency-domain information can be extracted both from the continuous and discrete-time signal. For the discrete time signal, DFT methods have been utilized for the frequency-domain information extraction.

To avoid the computational complexity and timing issues of the DFT we have utilized the Fast Fourier Transform (FFT) to extract the frequency-domain information.

As the Histopathological image contains three channels, the FFT coefficients have been extracted from each of the three channels:

$$\begin{cases} h_f^r = \text{FFT coefficient from red channel} \\ h_f^g = \text{FFT coefficient from green channel} \\ h_f^b = \text{FFT coefficient from blue channel.} \end{cases}$$

The first top "t" FFT coefficients have been selected from each of the channel where $t = h_1 \times h_2$:

$$\mathcal{H}_{DF} = \begin{cases} h_f^{rt} = \text{Top } t \text{ FFT coefficient from red channel} \\ h_f^{gt} = \text{Top } t \text{ FFT coefficient from green channel} \\ h_f^{bt} = \text{Top } t \text{ FFT coefficient from blue channel.} \end{cases}$$

Here \mathcal{H}_{DF} represent the feature matrix for the Case3a that is for CNN-DF.

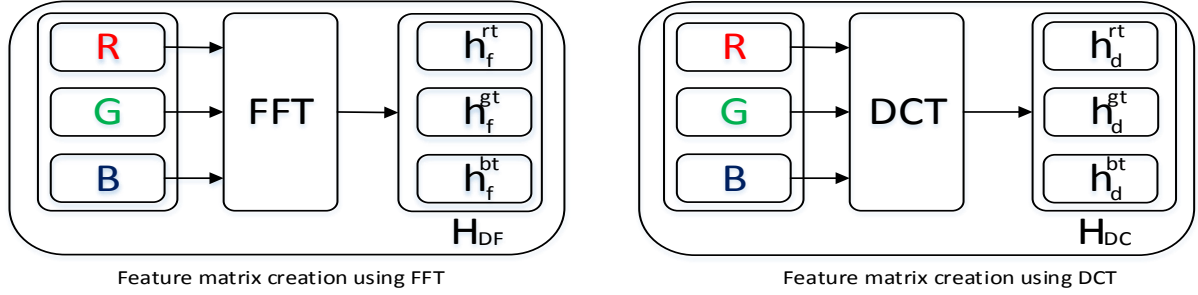


Figure 7.5: Feature-Selection Procedure from images when we use DFT DCT

DCT for feature selection

G. Strang first introduced the DCT method in 1974 [237]. A few DCT methods are available, and among them DCT-II methods have been largely utilized for image analysis. As a Histopathological image contains three channels, the DCT coefficients have been extracted from each of the three channels:

$$\begin{cases} h_d^r = \text{DCT coefficient from red channel} \\ h_d^g = \text{DCT coefficient from green channel} \\ h_d^b = \text{DCT coefficient from blue channel.} \end{cases}$$

The first top "t" FFT coefficients have been selected from each of the channels where $t = h_1 \times h_2$:

$$\mathcal{H}_{DC} = \begin{cases} h_d^{rt} = \text{Top } t \text{ DCT coefficient from red channel} \\ h_d^{gt} = \text{Top } t \text{ DCT coefficient from green channel} \\ h_d^{bt} = \text{Top } t \text{ DCT coefficient from blue channel.} \end{cases}$$

Here \mathcal{H}_{DC} represents the feature matrix for the Case3b that is CNN-DC.

Table 7.1 summarises extracted local features for different cases:

7.5 Convolutional Neural Network

A CNN model is a state-of-the-art method which has been largely utilized for image processing. A CNN model has the ability to extract global features in a hierarchical

Table 7.1: Number of Handcrafted Features

Case Name	CNN-CH	CNN-CL	CNN-DF	CNN-DC
Total Number of Features (Hand Crafted)	962	961	2883	2883

manner which ensures local connectivity as well as the weight-sharing property.

- **Convolutional Layer** The Convolutional layer is considered as the main working ingredient in a CNN model and plays a vital determining part of this model. A kernel (filter), which is basically a $n \times n$ matrix successively goes through all the pixels and extracts the information from them.

- **Stride and Padding**

The number of pixels a kernel will move in a step is determined by the stride size; conventionally the size of the stride keeps to 1. Figure 7.6 a shows an input data matrix of size 5×5 , which is scanned with a 3×3 kernel. The light-green image shows the output with stride size 1, and the green image represents the output with stride size 2. When we use a 3×3 kernel, and stride size 1, then the convolved output is a 3×3 matrix, however when we use stride size 2 the convolved output is 2×2 . Interestingly, if we use a 5×5 kernel on the above input matrix with stride 1 the output will be a 1×1 matrix. So the size of the output image has changed with both the size of the stride and the size of the kernel. To overcome this issue we can utilize extra rows and columns at the end of the matrices which contain 0s. This adding of rows and columns which contain only zero value is known as zero padding.

For example Figure 7.6 b shows how two extra rows have been added at the top as

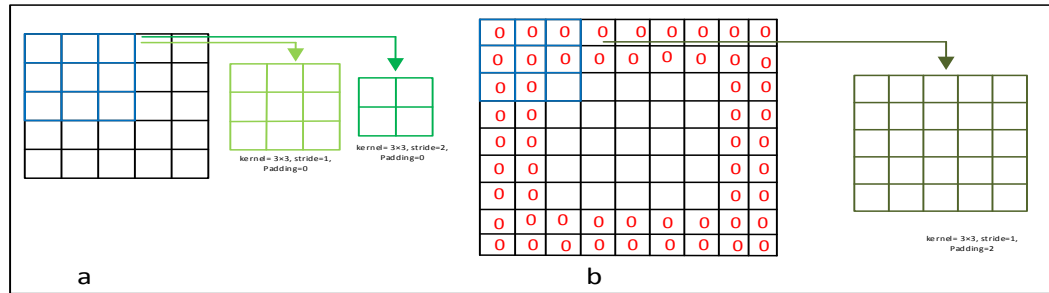


Figure 7.6: This figure represents the effects of kernel size, the size of stride and zero padding in a convolutional operation

well as the bottom of the original 5×5 matrix. Similarly, two extra columns have been added at the beginning as well as the end of the original 5×5 matrix. Now the olive-green image of Figure 7.6 b shows a convolved image where we have utilized a kernel of size 3×3 , stride size 1 and padding size zero. The convolved image is also a 5×5 matrix, which is the same as the original data size. So by adding the proper amount of zero padding we can reduce the loss of information which lies at the border.

- **Non-Linear performance**

Each layer of the NN produces linear output, and by definition adding two linear functions will also produce another linear output. Due to the linear nature of the output, adding more NN layers will show the same behavior as a single NN layer. To overcome this issue, a rectifier function such as Rectified Linear Unit (ReLU), Leaky ReLU, TanH, Sigmoid etc. has been introduced to make the output nonlinear.

- **Pooling Operation**

A CNN model produces a large amount of feature information. To reduce the feature dimensionality a down-sampling method named a pooling operation has been performed. A few pooling operation methods are well known such as

- Max Pooling
- Average Pooling.

For our analysis, we have utilized the Max Pooling operation which selects the maximum values within a particular patch.

• Drop-Out

Due to the over training of the model it shows very poor performance on the test dataset, which is known as over-fitting. This over-fitting issue has been controlled by removing some of the neurons from the network, which is known as Drop-Out.

• Decision Layer

For the classification decision, at the end of a CNN model a decision layer is introduced. Normally a Softmax layer or a SVM layer is introduced for this purpose. This layer contains a normalized exponential function and calculates the loss function for the data classification.

Figure 7.7 shows the work flow of a generalized CNN model which can be used for image classification. Before the decision layer, there must be at least one immediate dense layer available in a CNN model. Utilizing the softmax layer, the output of the last layer can be represented as

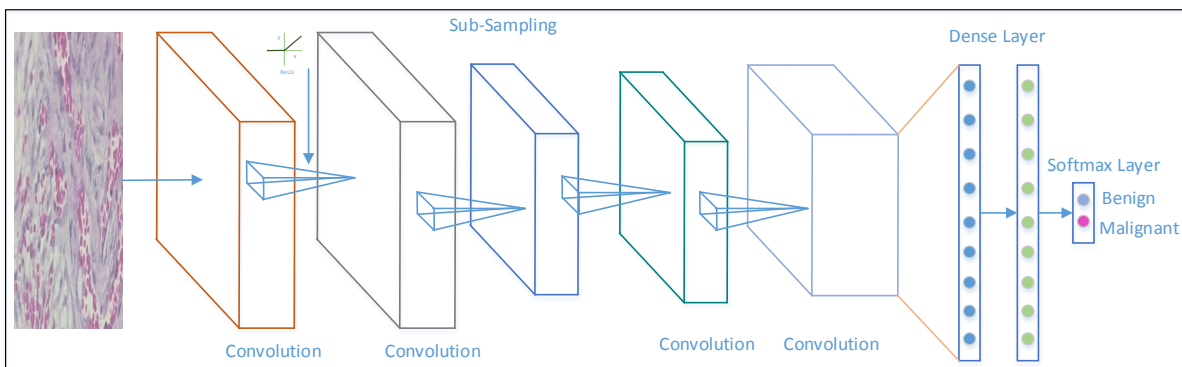


Figure 7.7: Workflow of a Convolutional Neural Network

$$\bar{\mathcal{Y}}_d = \frac{\exp(\mathcal{H}_d^{end})}{\sum_{d=1}^{class} \exp(\mathcal{H}_d^{end})} \quad (7.1)$$

where

$$\mathcal{Y}_d^{end} = \sigma(\mathcal{W}^{end} * \mathcal{H}_k^{end-1} + \mathcal{B}^{end}) \quad (7.2)$$

Here \mathcal{H}_k^{end-1} represents the k^{th} neuron at the $(end - 1)^{th}$ layer, and σ represents the nonlinear function. For binary classification the number of class = 2. Let $d = 1$ represent the Benign class and else it represents the Malignant class. The cross-entropy loss of $\bar{\mathcal{Y}}_d$ can be calculated as

$$\mathcal{L}_d = -\ln(\bar{\mathcal{Y}}_d). \quad (7.3)$$

As we are working on a two-class classification problem then only the \mathcal{L}_1 and \mathcal{L}_2 values are possible and the output will be Benign when $\mathcal{L}_1 \leq \mathcal{L}_2$ else the output will be Malignant.

7.5.1 CNN Model for Image Classification

For breast-image classification we have utilized the CNN model with the following architectures:

- Model-1: Model-1 utilizes a residual block, represented as Block-n. Each Block-n contains two convolutional blocks named C-n and R-n. The C-n layer convolves the input data with a 5×5 kernel along with a ReLU rectifier and produces 16 feature maps. The output $\mathcal{X}\mathcal{C}_n$ of the C-n layer passes through the R-n convolutional layer, which also utilizes a 5×5 kernel along with a ReLU rectifier. The R-n layer also produces 16 feature maps. The output $\mathcal{X}\mathcal{R}_n$ of the R-n layer is merged with the output $\mathcal{X}\mathcal{C}_n$ of the layer and produces a residual output. The output $\mathcal{X}\mathcal{R}_n$ of Block-n can be represented as

$$\mathcal{X}\mathcal{R}_n^1 = \sigma[\sigma(\mathcal{X}\mathcal{R}_n, \mathcal{W}_n + \mathcal{B}_n) + \mathcal{X}\mathcal{R}_n] \quad (7.4)$$

where \mathcal{W}_n represents the weight matrix and \mathcal{B}_n represents the bias vector.

The input matrix passes through Block-1 and Block-2 as shown in Figure 7.8 (left image). The output of Block-1 is fed to Block-3, the output of Block-3 is fed as an input to Block-5, the output of Block-5 is fed as an input to Block-7, the output of Block-7 is fed as an input to Block-9. Similarly the output of Block-2 is fed to Block-4, the output of Block-4 is fed as an input to Block-6, the output of Block-6 is fed as an input to Block-8, the output of Block-8 is feed as an input of the Block-10. Now the output of Block-9 and Block-10 are concatenated in the Concat layer. After the Concat layer a Flat Layer, a Drop-Out Layer and a Softmax layer have been placed one after another. The output of the Softmax layer has been used to classify the images into Benign and Malignant classes.

- Model-2: Model-2 utilizes almost the same architecture as Model-1. The only difference is that in each Block-n the output $\mathcal{X}\mathcal{C}_n$ of layer C-n is multiplied (rather than added) with the output $\mathcal{X}\mathcal{R}_n$ of layer R-n. The output of Block-n can be represented as

$$\mathcal{X}\mathcal{R}_n^2 = \sigma[\sigma(\mathcal{X}\mathcal{R}_n, \mathcal{W}_n + \mathcal{B}_n) \times \mathcal{X}\mathcal{R}_n] \quad (7.5)$$

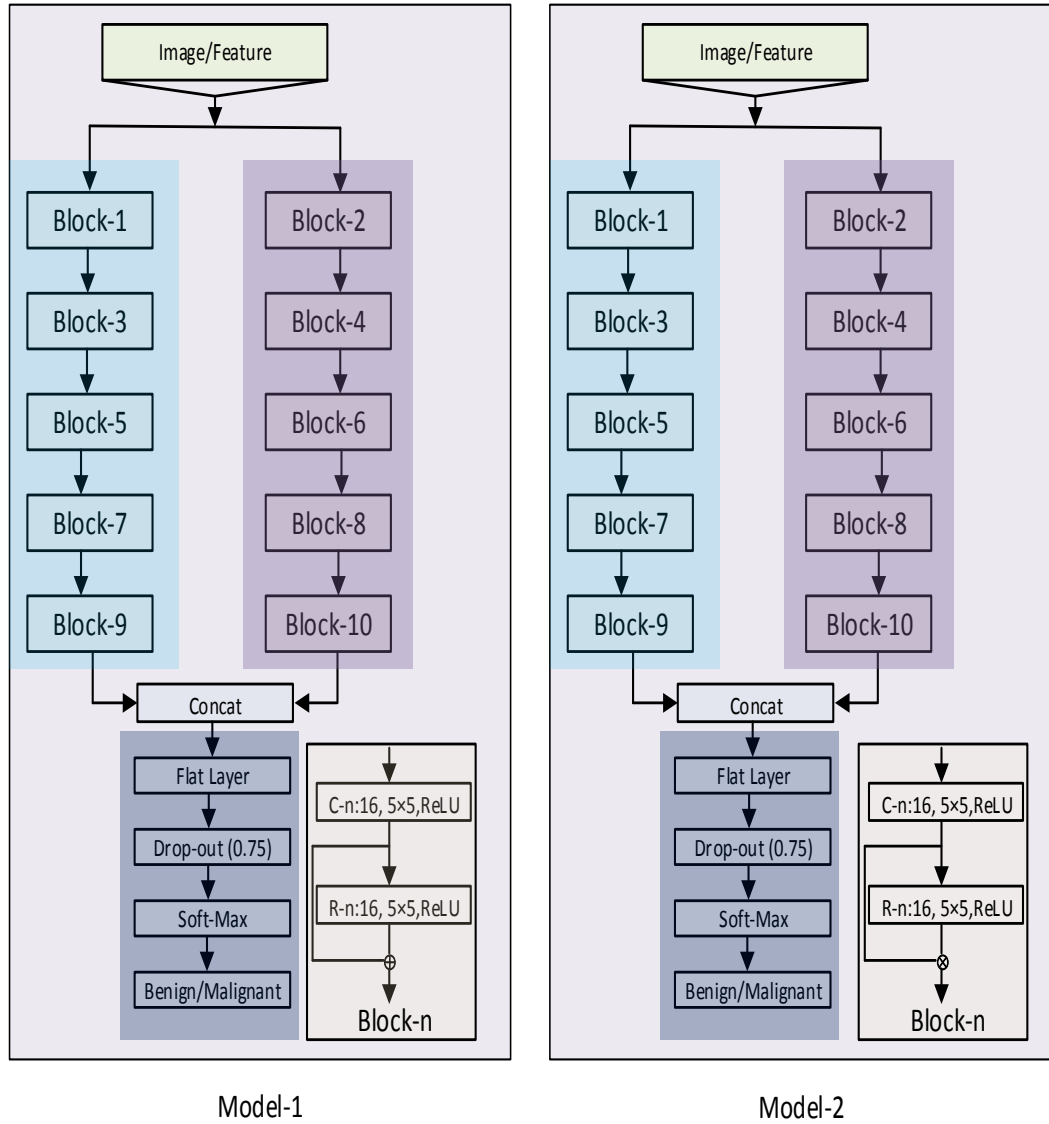


Figure 7.8: Architecture of Model-1 at the left and architecture of Model-2 at the right

7.6 Performance-Measuring Parameters and utilised Platform

The performance of a classifier is measured by some base mark criteria, which can be obtained by a two-dimensional matrix known as the Confusion Matrix [43]. The content of the matrix position $i = j$ represents how many times the target is correctly classified.

Table 7.2: A summary of classification-performance measurement parameters

Metric Name	Mathematical Expression	Highest Value	Lowest Value
Recall	$\frac{TP}{TP+FN}$	+1	0
Precision	$\frac{TP}{TP+FP}$	+1	0
Specificity	$\frac{TN}{TN+FP}$	+1	0
F-measure	$\frac{TP+TN}{TP+TN+FP+FN}$	+1	0
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	+1	-1

So it is expected that the non-diagonal positions of the Confusion Matrix should be as small as possible. Figure 7.9 shows a graphical representation of a Confusion Matrix and Table 7.2 summarizes a few of the well-known classification performance measurement parameters.

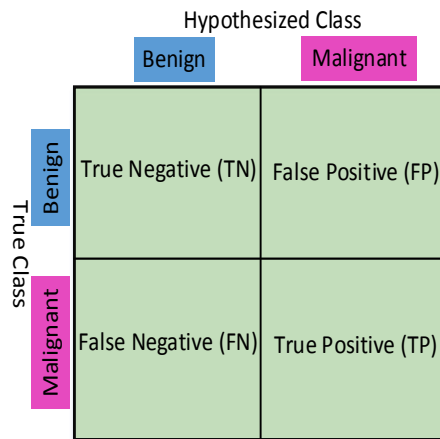


Figure 7.9: Confusion Matrix

Platforms Used

Image pre-processing related tasks are performed in MATLAB@16. Out of the available platforms for CNN model development, we have selected Keras. Lastly, most of the matrix operations are performed on a GeForce GTX 1080 GPU, as the classification of images involves billions of matrix operations, which is not possible with a low-grade CPU.

7.7 Results and Discussion

For the classification, we utilized the BreakHis data set [205]. The images of this dataset are RGB in nature, having 8-bit depth and a PNG extension. The images are 700×460 pixels in size. All the images are divided into four groups, depending on the visual magnification factor, namely $40\times$, $100\times$, $200\times$ and $400\times$, where \times represents the magnification factor. We performed our experiments on the individual groups of the dataset.

7.7.1 Performance of $40\times$ Dataset

Table 7.3 shows the performance of Model-1 and Model-2 on the $40\times$ data-set. The overall best performance is achieved when CNN-CH along with Model-1 is utilized. In this situation the achieved Accuracy is 94.40%, where the Recall and Precision values are 96.00% and 86.00% respectively. For the Model-1, CNN-CL provides a similar performance. When we use Model-1 the worst Accuracy of 86.47% is achieved when we utilize CNN-DC.

For Model-2 the best Accuracy of 88.31% is achieved when we utilize the CNN-I algorithm. However the achieved Recall value is 96.00% and the Specificity value is 69.45%, which indicates that almost 31.00% of the Benign images have been mis-classified as Malignant images. When we utilize the CNN-CH algorithm along with Model-2, the Recall value is 100.00% and FPR is 100.00%. This indicates that all the data, irrespective

Table 7.3: Performance of various cases on $40\times$ dataset

		Accuracy %	TNR/ Specificity %	FPR (%)	FNR (%)	TPR/ Recall (%)	Precision (%)	F-Measure (%)
Model-1	CNN-CH	94.40	86.00	14.00	04.00	96.00	94.00	95.00
	CNN-CL	93.32	85.05	15.95	03.20	96.70	94.00	95.00
	CNN-I	87.47	79.31	20.68	10.00	91.00	91.00	91.00
	CNN-DF	88.31	78.16	21.80	07.52	92.47	91.00	92.00
	CNN-DC	86.47	74.37	25.86	08.47	91.52	90.00	91.00
Model-2	CNN-CH	70.00	0.00	100.00	0.00	100.00	66.00	80.00
	CNN-CL	80.30	45.77	54.07	5.6	94.35	81.00	79.00
	CNN-I	88.31	69.45	30.45	4.00	96.00	88.50	84.78
	CNN-DF	85.47	73.56	26.43	9.40	90.35	89.30	83.45
	CNN-DC	86.50	52.87	47.12	3.2	96.72	83.00	90.00

of Benign or Malignant, are classified as Malignant. In terms of Accuracy, CNN-DF and CNN-DC provide a similar performance, however CNN-DF provides better specificity performance than CNN-DC. More specifically, CNN-DC mis-classifies almost 50.00% of the Benign data as Malignant data.

Figure 7.10 a, b and c represents the Train and Test Accuracy, loss, M.C.C. values when we utilized Model-1 and CNN-CH on the $40\times$ dataset. Up to around epoch 25, the Train Accuracy and Test Accuracy remained almost the same; after around the 25th, the Train Accuracy rapidly increased but the Test Accuracy increased very slowly. As the epoch proceeds the Train Accuracy remains almost constant. For the loss performance, after around epoch 25, the Train loss continues to decrease, however the Test loss increases. The loss difference between the Train and Test Loss continuously increases as the epoch proceeds. For this case the M.C.C. value is never negative. Up-to around epoch 25, Train and Test M.C.C. values remain almost constant. After 25 epoch the train M.C.C. values continuously improve but the test M.C.C. values remained constant around 00.86.

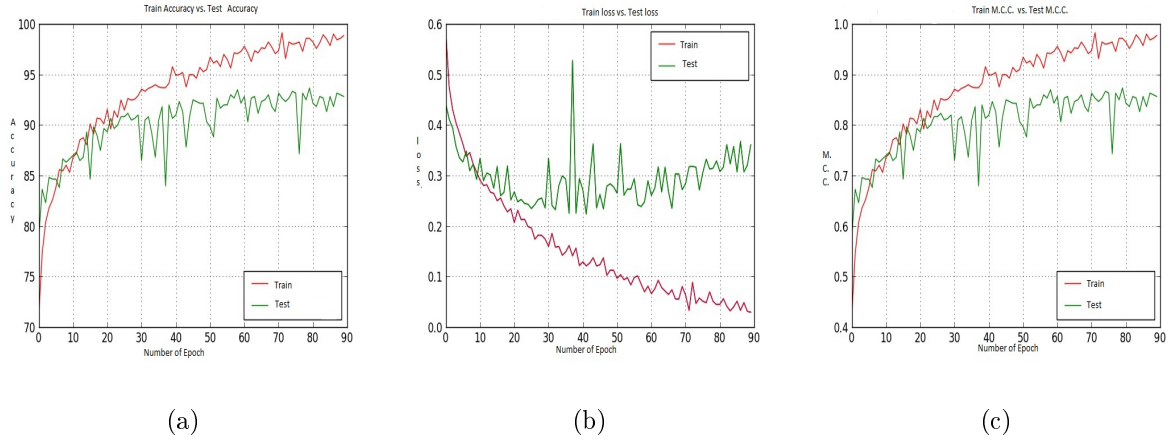


Figure 7.10: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the $40\times$ dataset.

Figure 7.11 shows the Accuracy, loss and M.C.C values for Model-2 on the $40\times$ dataset.

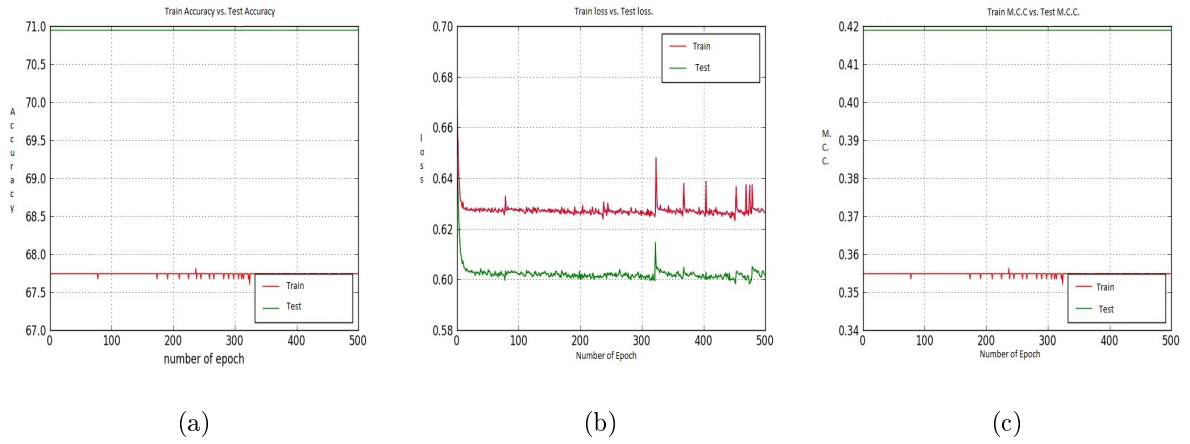


Figure 7.11: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the $40\times$ dataset.

Among all the available Models and Cases CNN-CH provides the worst performance on the $40\times$ dataset when we utilize Model-2. In this particular situation, the Train Accuracy (71.00%) and the Test Accuracy (64.00%) are constant throughout the epochs. For the loss performance the loss values for Train and Test are also constant throughout the epochs. When we utilised Model-1 and the CNN-CH algorithm on the $40\times$ dataset we

ran our experiment only until about epoch 90 and got quite constant performance.

7.7.2 Performance of 100× Dataset

For the 100× dataset when we utilized Model-1 and the CNN-CH algorithm it provides almost 95.93% Accuracy, along with 94.85% Specificity and 96.36% Recall values. This indicates that only 5.15% of the Benign data has been mis-classified as Malignant data, and 3.64% of Malignant images have been mis-classified as Benign images. When we use CNN-I, that is when we utilized

Table 7.4: Performance of various cases on 100× dataset

		Accuracy %	TNR/ Specificity %	FPR (%)	FNR (%)	TPR/ Recall (%)	Precision (%)	F-Measure (%)
Model-1	CNN-CH	95.93	94.85	05.15	03.64	96.36	98.00	97.00
	CNN-CL	92.00	89.10	10.90	06.70	93.30	96.00	94.00
	CNN-I	87.15	67.42	32.50	05.00	95.00	88.00	95.00
	CNN-DF	89.26	81.14	18.85	07.50	92.5	93.00	93.00
	CNN-DC	87.15	78.28	21.71	09.31	90.68	91.00	91.00
Model-2	CNN-CH	67.96	43.00	57.00	22.00	78.00	78.00	78.00
	CNN-CL	78.53	31.42	68.52	2.73	97.27	78.00	75.00
	CNN-I	86.12	81.87	18.18	11.26	88.78	93.00	87.00
	CNN-DF	85.47	85.71	14.20	17.95	82.05	94.00	87.00
	CNN-DC	86.11	65.71	34.28	6.13	93.86	87.00	90.00

raw images as input with Model-1, the Accuracy is 87.15%. In this particular situation, the Recall value is 93.30% but the Specificity value is 67.42%, This indicates that almost one third of the Benign images have been mis-classified as Malignant images, and this low Specificity value reduces the overall performance. CNN-DC and CNN-DF show similar performance when we utilized Model-1 and the 100×. For Model-2, when we utilized CNN-I, that is when we utilized raw images as input, it produces the best accuracy

among all the cases. In this particular case the Specificity value is 81.87% and the Recall value is 88.78%. CNN-DC also provides similar Accuracy to CNN-I , however it shows very poor specificity performance of 65.71% . For Model-2 CNN-CH provides the worst performance among all the available cases with 67.96% accuracy, 43.00% Specificity and 78.00% Recall values.

Figure 7.12 shows the Accuracy, Loss and MCC values for the CNN-CH case on the $100\times$ dataset when Model-1 has been utilized. Initially upto around epoch 25, the Test Accuracy values show better performance than the Train Accuracy. After that Train Accuracy shows better performance than Test Accuracy. After around epoch 50 Test Accuracy is about 96.00%, and Test Accuracy is about 95.00%. For the loss performance, up to around epoch 21, the Test loss shows better values than the Train loss. However after epoch 21, the Train loss continuously decreases whereas the Test loss shows poor performance. For the M.C.C values, after around epoch 80 the Train M.C.C value is 0.98 and the Test M.C.C value is around 0.95.

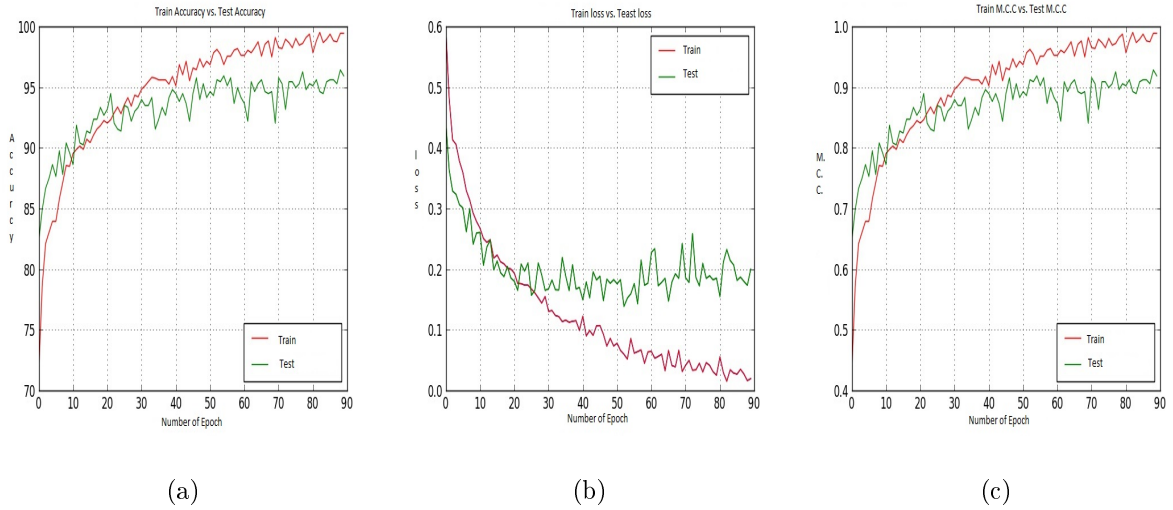


Figure 7.12: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the $100\times$ dataset.

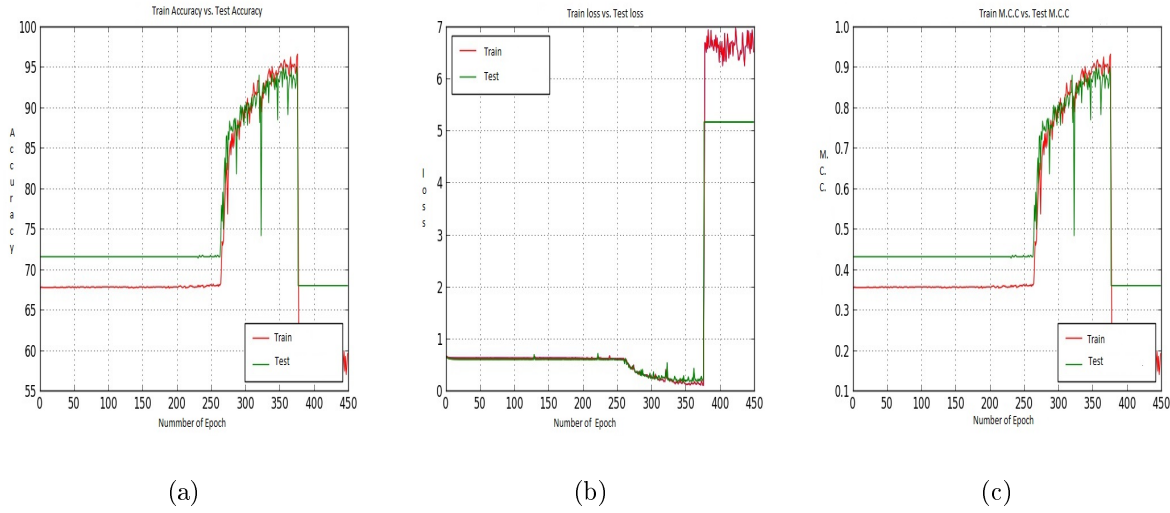


Figure 7.13: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the $100\times$ dataset.

7.7.3 Performance of $200\times$ Dataset

For the $200\times$ dataset, when Model-1 and CNN-CH are used together 97.90% Accuracy is achieved, along with 94.94% Specificity and 98.20% Recall values. This indicates that almost all the Malignant data has been classified into Malignant; whereas 5.06% of Benign data has been mis-classified as Malignant.

When we use Model-2 along with CNN-CH, CNN-CL or CNN-DC, on the $200\times$ dataset, we get very poor performance. In all these three cases, all the data is classified as Malignant data irrespective of reality. In this scenario, the best performance is achieved when we utilized the raw image as input, that means the CNN-I case. In this case, we achieved 86.00% Accuracy, along with 81.87% Specificity and 88.78% Recall values.

Figure 7.14 shows the Accuracy values for the CNN-CH case on the $200\times$ dataset. Up to around epoch 15 Train Accuracy show almost the same performance with some exceptions. After that Train accuracy shows slightly better performance than Test accuracy. Train data shows 100% Accuracy around epoch 90, whereas Test Accuracy shows 97.00%.

Table 7.5: Performance of various cases on $200\times$ dataset

		Accuracy %	TNR/ Specificity %	FPR (%)	FNR (%)	TPR/ Recall (%)	Precision (%)	F-Measure (%)
Model-1	CNN-CH	97.19	94.94	5.06	1.70	98.20	98.00	98.00
	CNN-CL	94.00	92.42	07.57	05.65	09.41	96.00	95.00
	CNN-I	86.44	79.31	24.74	08.10	91.89	88.00	86.00
	CNN-DF	87.10	88.38	11.60	13.51	86.48	94.00	90.00
	CNN-DC	85.61	71.71	28.28	08.00	92.00	87.00	90.00
Model-2	CNN-CH	67.60	1.00	98.98	0.00	100.00	67.00	81.00
	CNN-CL	67.27	0.00	100.00	0.00	100.00	67.00	80.00
	CNN-I	86.00	81.87	18.18	11.26	88.78	93.00	87.00
	CNN-DF	85.28	72.22	27.77	8.30	96.60	87.00	89.00
	CNN-DC	67.00	0	100.00	0	100.00	67.00	80.00

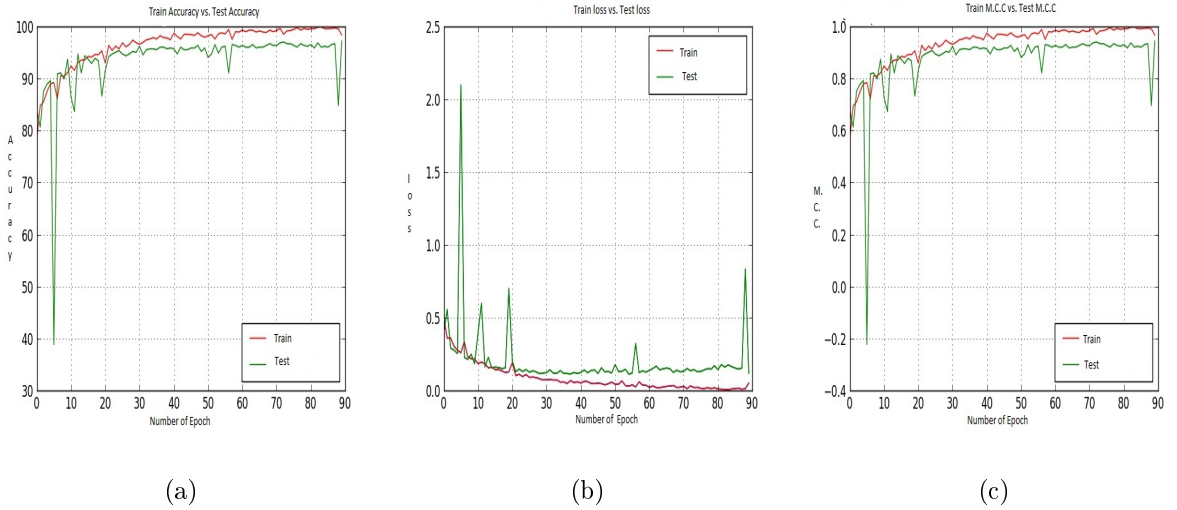


Figure 7.14: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the $200\times$ dataset.

For the loss performance, as shown in Figure 7.14 (b), the Train loss is almost 0, whereas Test loss also shows quite small values but not 0. After around epoch 20, the Train and Test loss values remained constant. After that the difference between the Train loss value and the Test Loss value is constant. As the epoch proceeds the Train and Test M.C.C

values increased. Around epoch 85 the Train M.C.C value touches the highest M.C.C value where as the Test M.C.C value is around 0.91.

We saw earlier using CNN-CH, CNN-CL and CNN-DC with Model-2 on the $200\times$ dataset gave very poor performance. Figure 7.15 shows the Accuracy, Loss and MCC values when we utilized the CNN-CH algorithm on the $200\times$ dataset. For the Accuracy case, Train shows around 69.00% accuracy for all epochs up to 450. On the other hand Test accuracy remains constant at around 67.5% through out the epochs. For the loss performance the difference between the Train and Test losses remains the same throughout all the epochs.

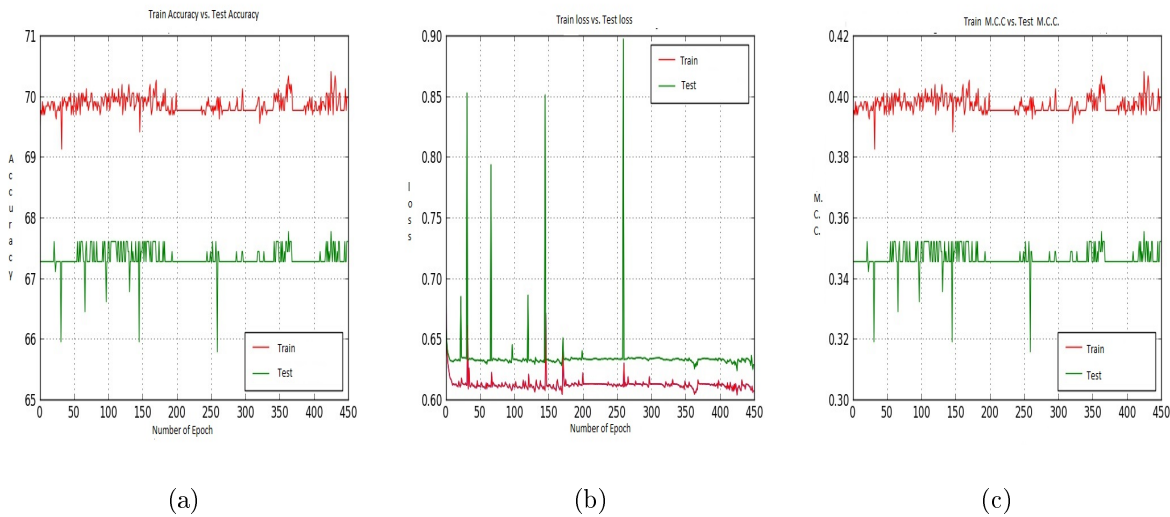


Figure 7.15: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH on the $200\times$ dataset.

7.7.4 Performance of $400\times$ Dataset

When we use Model-1 and the CNN-CH algorithm on the $400\times$ dataset, the best performance is achieved. In this case the Accuracy is 96.00%, with 90.16% Specificity and 97.79% Recall values. CNN-DF and CNN-DC provide similar performance. When we utilized raw images as an input, the Accuracy achieved is 84.43%. When we use Model-2 and

the CNN-CH algorithm on the $400\times$ dataset, the system gives the worst performance, of 67.80% Accuracy with 0.005% Specificity and 100.00% Recall values. Interestingly, CNN-I, CNN-DF, and CNN-DC provide similar performance.

Table 7.6: Performance of various cases on $400\times$ dataset

		Accuracy %	TNR/ Specificity %	FPR (%)	FNR (%)	TPR/ Recall (%)	Precision (%)	F-Measure (%)
Model-1	CNN-CH	96.00	90.16	9.84	2.2	97.79	95.00	96.00
	CNN-CL	80.00	80.87	19.10	6.60	9.39	91.00	92.00
	CNN-I	84.43	70.49	29.50	8.50	91.46	86.00	89.00
	CNN-DF	93.00	87.43	12.56	7.70	92.30	94.00	93.00
	CNN-DC	92.00	85.70	14.20	7.10	92.20	93.00	93.00
Model-2	CNN-CH	67.80	0.005	99.95	0.00	1.00	0.67	0.80
	CNN-CL	66.48	00.00	100.00	0.00	100/00	44.00	53.00
	CNN-I	86.34	75.40	24.59	10.46	89.50	88.00	89.00
	CNN-DF	87.17	74.86	25.13	6.61	93.30	88.00	91.00
	CNN-DC	86.26	73.22	26.77	7.11	92.83	87.00	90.00

Figure 7.16 shows the Accuracy, Loss and M.C.C values for different epochs when we utilized Model-1, CNN-CH and $400\times$ dataset. Figure 7.16 shows that, up to around epoch 15, the Train and Test Accuracy, Loss and MCC values remain almost the same with some exceptions. After around epoch 15, Train accuracy shows better performance than Test accuracy. After epoch 50 the Train accuracy become constant at around 96.00% whereas the Test accuracy shows continually better performance. For the loss, as the epoch proceeds the difference between the Train loss and the Test loss increased. For M.C.C, the Train M.C.C value touches around 0.92.

Figure 7.17 shows Accuracy, Loss and M.C.C values for different epochs when we utilized Model-2, CNN-CH and the $400\times$ dataset. In this particular scenario, the Train Accuracy keeps around 68.25% whereas the Test Accuracy is around 66.50%. For the loss

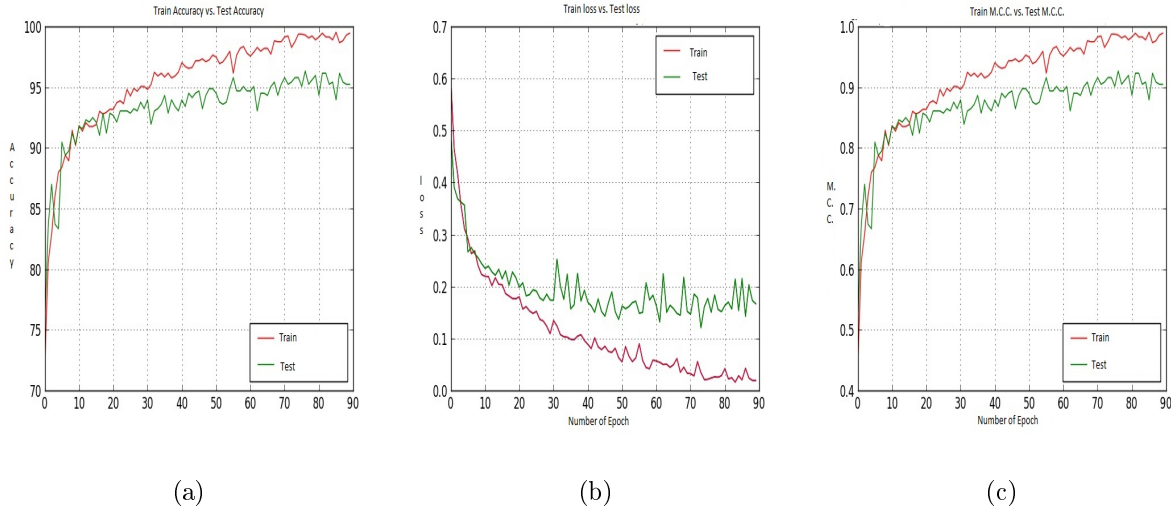


Figure 7.16: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-1, CNN-CH on the $400\times$ dataset.

case, the Train loss remained around 0.625 and the Test loss remained around 0.63; with some exceptions those values remain constant for all epochs.

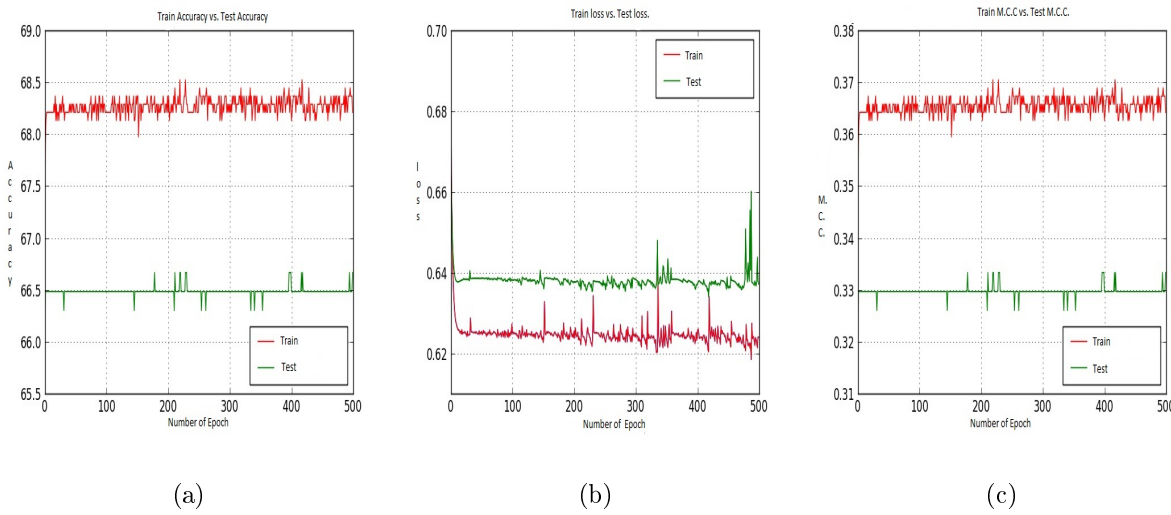


Figure 7.17: (a), (b), and (c) represent the Train and Test Accuracy, Loss and M.C.C. values when we utilise Model-2, CNN-CH case on the $400\times$ dataset.

7.7.5 Required Time and Parameters

Table 7.7 shows the number of parameters required and the time required to run per epoch for Model-1 and Model-2. Model-1 requires 119666 parameters for the total operation whereas Model-2 requires 120466 parameters.

Table 7.7: Required Time and Number of Parameters

Model	Case	Parameters	Time (s)	Model	Case	Parameters	Time (s)
Model-1	CNN-CH	120466	45	Model-2	CNN-CH	119666	45
	CNN-CL	119666	45		CNN-CL	120466	45
	CNN-I	119666	38		CNN-I	120466	38
	CNN-DF	119666	40		CNN-DF	120466	40
	CNN-DC	119666	40		CNN-DC	120466	40

7.7.6 Comparison with Findings

Table 7.8 summarizes a few recent findings of Histopathological breast-image classification. Brook et al. [238] utilize a total of 361 images for their experiment. From each of the images they have collected 1060 features and as a classifier tool they have utilized the SVM method and obtained 96.40% accuracy. Zhang et al. [239] also perform the classification operation on the same dataset utilizing ensemble methods and have obtained 97.00% Accuracy. Ciresan et al. [240] and Wang et al. [241] both perform their experiments on the ICPR12 dataset where they have utilized global features. The finding of our chapter cannot be compared directly with the above-mentioned findings because they have performed their experiment on a different dataset as well as using different classification techniques.

Spanhol et al. [7], Han et al. [242] and Dimitropoulos et al. [216] perform their experiment on the BreakHis dataset. Spanhol et al. [7] obtained best performance when they uti-

lized the $40\times$ dataset and obtained 90.40% accuracy. Han et al. [242] achieve $95.80\pm3.10\%$ Accuracy on the $40\times$ dataset. Dimitropoulos et al. [216] obtained best Accuracy performance when they utilized the $100\times$ dataset and the VLAD method. Our experiment has been performed on the BreakkHis dataset, and obtained best Accuracy 97.19%, which is almost comparable with the the state-of-the-art findings of Han et al. [242]. Both the work by Spanhol et al. [7], Han et al. [242] and Dimitropoulos et al. [216] finds the Accuracy values only, however in this chapter we have also findings for the Specificity, Precision, Recall values along with finding the required number of parameters and the time required to perform the experiment. Beside this we have also compared our result with Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) methods and found that our algorithms provide better performance than those two methods

Table 7.8: Summarises a few recent findings of Histopathological breast-image classification

Authors	Dataset Details	Features	Classification Tool	Number of classes	Accuracy %	Sensitivity %	Recall %	Precision %	ROC %
Brook et al. [238]	Total Sample: 361	1. Local Features 2. 1050 Features.	SVM	1. 3 Classes					
				a. normal tissue	96.40	—	—	—	—
				b. carcinoma as situ					
Zhang [239]	Total Sample: 361	1. Local Feature 2. Textural property 3. Curvelet transform	Ensemble	c. invasive ductal					
				1. 3 Classes					
				a. normal tissue	97.00	—	—	—	—
Ciresan et al. [240]	ICPR12	1. Global Features	DNN	b. carcinoma as situ					
Wang et al. [241]	ICPR12	1. Global Feature	—	c. invasive ductal					
		2. Textural Features							73.45
Wang et al. [243]	Total 68 images	1. Local Features	SVM		95.50	99.32	94.14	—	—
	BreakHis								
	a. 40×				a. 90.40				
	b. 100×	Global Features	CNN		b. 87.40	—	—	—	—
Spanhol et al. [7]	c. 200×				c. 85.00				
	d. 400×				d. 83.00				
Han et al. [242]	BreakHis								
	a. 40×				a. 95.80±3.1				
	b. 100×	Global Features	VLAD		b. 96.90±1.9	—	—	—	—
	c. 200×				c. 96.70±2.0				
Dimitropoulos et al. [216]	d. 400×				d. 94.90±2.8				
	BreakHis								
	a. 40×				a. 91.80				
	b. 100×	Global Features	VLAD		b. 92.10	—	—	—	—
	c. 200×				c. 91.40				
	d. 400×				d. 90.20				

7.8 Conclusion

This chapter has classified a set of Histopathological breast-images into Benign and Malignant classes. For the classification, the state-of-the-art CNN model has been utilized along with residual blocks. The CNN method generally extracts global features while maintaining the hierarchical structure. However local features and frequency domain information also carry significant information from images which help significantly for the image classification. Utilizing the benefit of local and frequency domain information as well as hierarchical property of the CNN model this chapter has proposed two different sets of algorithms. The first set of algorithms extracts local feature information whereas the second set of algorithms extracts frequency-domain information. Feature-extraction based cases suggested two distinct algorithms, where the first algorithm utilized the Contourlet Transform as well as Histogram-based information, whereas the second algorithm is based on the Contourlet Transform and Local Binary Pattern information. Frequency-feature based cases also provide two algorithms, one of the algorithms based on DFT-based information whereas the second is based on the DCT algorithm. This chapter has utilized the BreakHis dataset for the experiment, which contains four datasets. Most of the recent findings on this dataset analyze the Accuracy information. In this chapter, along with finding the Accuracy information, we have also found the Precision, Recall, Specificity, MCC and F-1 Score values. Experiment shows that, the CNN-CH case provides the best performance on all the available datasets. Specifically, the $200\times$ dataset provides the best performance of the available datasets with 97.19% Accuracy, 94.94% Specificity and 98.20% Recall value. The computational complexity and required time for the classification are two important parameters for the CNN-based image-classification task. In this chapter we have investigated how many parameters are required and the time required for the experiment, which provides information about the complexity of this technique.

Chapter 8

Histopathological Breast-Image Classification with Restricted Boltzmann Machine along with Back Propagation

8.1 Abstract

Deaths due to cancer have increased rapidly in recent years. Among all the cancer diseases, breast cancer causes many deaths in women. A digital medical photography technique has been used for the detection of breast cancer by physicians and doctors, however, they need to give more attention and spend more time to reliably detect the cancer information from the images. Doctors are heavily reliant upon Computer Aided Diagnosis

Accepted as: A. A. Nahid, A. Mikaelian and Y. Kong, "Histopathological Breast Image Classification with Restricted Boltzmann Machine along with Back Propagation", *BioMed Research, Allied Academics*. [Accepted on 2nd April, 2018].

(CAD) for cancer detection and monitoring of cancer. Because of the dependence on CAD for cancer diagnosis, researchers always pay extra attention to designing an automatic CAD system for the identification and monitoring of cancer. Various methods have been used for the Breast-Cancer image-classification task, however, state-of-the-art Deep Learning techniques have been utilised for cancer image classification with success due to its self-learning and hierarchical Feature-Extraction ability. In this chapter we have developed a Deep Neural Network (DNN) model utilising a Restricted Boltzmann Machine with "Scaled Conjugate Gradient" back propagation to classify a set of Histopathological Breast-Cancer images. Our experiments have been conducted on the Histopathological images collected from the BreakHis dataset.

8.2 Introduction

Many patients in the world suffer from cancer. There are different kinds of cancer, among them Breast Cancer (BC) is a prominent one, and is specifically a serious health threat to women. As a case study, Figure 8.1 shows the death statistics due to BC in Australia for the last 5 years. This figure shows that the death trend due to BC increased every year at an alarming rate in Australia. This might be considered as an example of the BC situation throughout the world. Obviously this causes a serious human and social impact. Proper and timely detection of BC can save or at least improve the condition of susceptible people. Along with other conditions, the detection of BC largely depends on investigation of biomedical images captured by different imaging techniques such as X-Rays, Mammogram, Magnetic Resonance, histopathological images, etc. For perfect diagnosis of BC, a biopsy can produce reliable results with confidence. Histopathological images are used as a standard image for cancer diagnosis. However, their analysis is very time-consuming and needs extra attention for the perfect diagnosis along with the

expertise of the physicians and doctors.

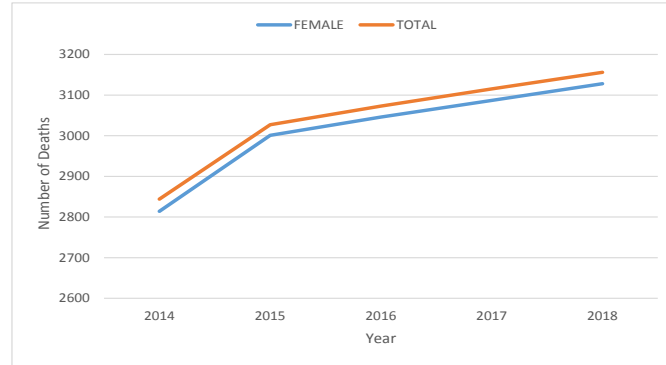


Figure 8.1: Death statistics due to BC for the last 5 years in Australia

The history of using Machine-Learning techniques for general image classification is a long one. Using the advancement and the deliverable engineering of image classification, scientists have used such techniques for medical image classification. An important part of the image classification is appropriate selection of features such as the Gray-Level Co-occurrence Matrix (GLCM), Tamura, etc. as well as classifier models such as Support Vector Machine (SVM), Random Tree (RT), Random Forest (RF), etc [244]. In a few cancer image-classification cases, scientists also extract information on nuclei. J. Diz et al. utilised both GLCM and Gray-Level Run Length Matrix (GLRLM) for mammogram image (400 images) classification and achieved 76.00% accuracy [113] where they employed the RF algorithm. The RF algorithm has also been used for histopathological image classification. Y. Zhang et al. [117], D. Bruno et al. [120], and A. Paul et al. [115] utilised histopathological images with different features. A. Paul et al. [115] utilised the Haralick features, D. Bruno et al. [120] used the Curvelet transform and Local Binary Pattern (LBP), Y. Zhang et al. [117] implemented the Curvelet transform, GLCM and CLBP together for classification.

The SVM is another popular and useful classifier for image classification. For the very first time A. Bazzani et al. utilised SVM techniques for breast image classification. L.O.

Martins et al. [136] utilised Ripley's K Function along with an SVM for Mammogram image classification and obtained accuracy, sensitivity and specificity of 94.94%, 92.86% and 93.33%, respectively. R.F. Chang et al. [133] utilised an auto-correlation coefficient for ultrasound breast-image classification and obtained 85.6% accuracy. S. Kavitha et al. [143] implemented histogram, textural (using the Gabor Filter) features and a few clinical features which were extracted from the images. They also resorted to SVM techniques for the image classification and obtained 90% Accuracy. R. Chang et al. classified a set of tomography images (250 images) using SVM techniques where the images are surrounded by speckle noise [125]. Fractional Fourier Transform (FFT) information has been used as features by Y.D. Zhang et al. [137] for Mammogram image classification using SVM along with Principal Component Analysis (PCA) techniques. J. Dheba et al. [140] utilised Laws texture features to classify images into Benign and Malignant (MIAS database) and achieved 86.10% accuracy. They performed their experiment on 200 images and obtained $92.16 \pm 3.60\%$ accuracy. It is found that the kernel method along with the SVM technique can improve the classifier performance. I Naga et al. [124] classified the Micro-calcification clusters in Mammogram images using Gaussian and polynomial kernels.

Along with other classifier techniques NN techniques have always been a strong tool for image classification. In 1991 A. Dawson et al. utilised an NN for a BC image classifier [49]. Literature shows that, the Neural Network technique has been very successful for the analysis and classification of images. Recently, the Deep Neural Network (DNN) technique has emerged as a popular method for the analysis of images for the classification task, following the famous model AlexNet proposed by Alex Krizhevsky et al. [219]. They proposed their techniques for the image-classification issues [219] based on a Convolutional Neural Network (CNN), a branch of DNN. After the work of Alex Krizhevsky, advanced engineering of this technique has been used for various image-classification tasks. Hai Su et al. proposed a fast-scanning Deep Neural Network (fCNN) method for the image-

classification task [245], where they utilised seven convolutional layers for analysis of the images. Y. Wu et al. [80] used CNN for Global Feature-Extraction for Mammogram (40 images) image classification and achieved a Sensitivity of 75.00% and Specificity 75.00%. Mammographic breast-density classification was done using HT-L3 convolution by P. Fonseca et al. [87]. H. Rezaeilouyeh et al. [7] implemented both local and global features and utilised CNN for histopathological image classification. They utilised the Shearlet Transform for extracting local features and achieved a best Accuracy of $86 \pm 3.00\%$. J. Xu et al. [225] utilised the DCNN-Ncut-SVM methods together for Histopathological breast-image classification and obtained an ROC of 93.16%. For Nuclease detection, the Spatially Constrained CNN was employed by K. Sirinkunwattana et al. [246]. B. Huynh et al. combined transfer learning and ensemble techniques for Mammographic image classification. T. Kooi et al. [103] resorted to global crafted features along with the Transfer learning method (VGG model) for Mammographic image classification.

The Deep Belief Network (DBN) is another branch of DNN which is a recent concept, proposed by Hinton et al. in 2006 [247]. For the first time they used Restricted Boltzmann Machine (RBM) techniques for Modified National Institute of Standards and Technology (MNIST) character recognition. Discriminative Deep Belief Networks (DDBN) were proposed by Yan Liu et al. for visual data classification and they utilised backpropagation techniques [89]. Ahmed et al. preferred the DBN method for the breast-cancer classification task [248]. For their analysis, they used the Wisconsin Breast Cancer Data set, which gives nine features for each image. So, instead of directly working on the images, the authors used the available features and DBN techniques with backpropagation.

The literature shows that a few studies have been performed on Histopathological breast-image classification using Tamura features. Most of the work has been conducted on well-known datasets like MIAS and DDSM along with some Histopathological images. Fabio A. Shanol et al. provide a new set of Histopathological breast images in the BreakHis

dataset and they did BC image classification using a few different classifiers

Doing image classification largely relies on how we select the features for the classification task. In this chapter we have classified Histopathological (BreakHis) breast images using Tamura features and RBM along with contrast corrections. The overall architecture of this chapter is organised as follows: Section 8.2 gives a brief description concerning the Breast-Image classification issues; Section 8.3 image-classification model; Section 8.4 describes the proposed RBM model for the classification; Section 8.5 describes the contrast correction algorithms in a brief; Section 8.6 describes the Feature-Extraction methodology; Section 8.7 describes and analyses the results; and Section 8.8 concludes the chapter.

8.3 Image-Classification Model

Successful image classification depends on a number of steps such as image pre-processing, Feature-Extraction and using image-classifier tools. Depending on the image pre-processing steps we have proposed two algorithms:

- Algorithm-1: This algorithm does not apply any pre-processing steps before Feature-Extraction. Algorithm-1 directly extracts Tamura features from each image, and the features are fed to the proposed model of the Restricted Boltzmann Machine (RBM) for image classification. Figure 8.2 shows the overall workflow of Algorithm-1.

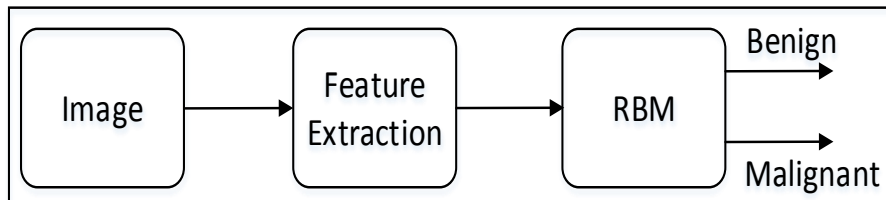


Figure 8.2: Workflow of Algorithm-1

- Algorithm-2: In the pre-processing steps, this algorithm enhances the contrast of each image in the dataset using the proposed contrast-enhancement algorithm, and then extracts the features. After that all the features are fed to the proposed model of the Restricted Boltzmann Machine (RBM) for image classification. Figure 8.3 shows the overall workflow of Algorithm-2.

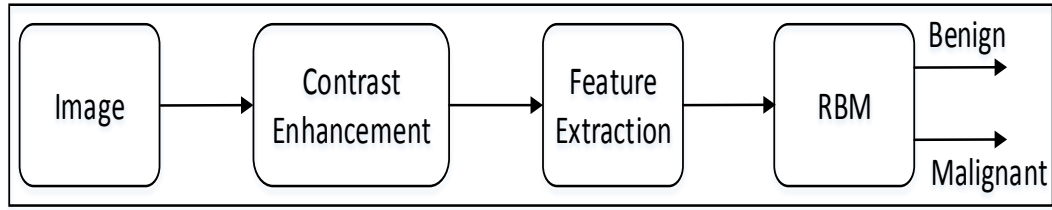


Figure 8.3: Workflow of Algorithm-2

8.4 Proposed RBM Model for Image Classification

In 1985, G. Hinton et al. proposed a Boltzmann machine (BM), which contains two layers named visible and hidden. The Restricted Boltzmann Machine (RBM) uses the concept of the BM. The difference between the RBM and the BM is that the connections of the hidden and visible layers are disjointed in an RBM. That is, in an RBM there are no intra-connections between the hidden layers and the visible layers. Figure 8.4 illustrates the BM and RBM machines. Let v and h represent the set of visible and hidden units. The energy of the joint configuration $\{v, h\}$ for BM can be defined as [249], [247]

$$E(v, h) = -\frac{1}{2}v^T L v - \frac{1}{2}h^T J h - v^T W h + \text{Bias} \quad (8.1)$$

where

- W is the weight between the visible and the hidden layers.
- L is the weight from visible to visible layer.

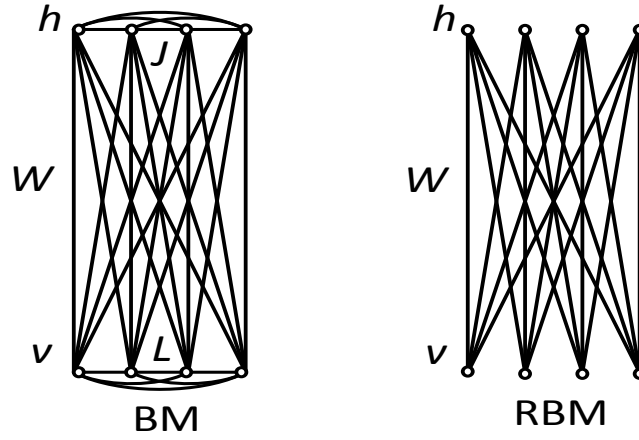


Figure 8.4: Graphical representation of BM and RBM models

- J is the weight from hidden layer to hidden layer.

Since we are working on an RBM, therefore $L = J = 0$. So we have

$$E(v, h) = -v^T W h + \text{Bias} \quad (8.2)$$

$$E(v, h) = -v^T W h - a^T v - b^T h \quad (8.3)$$

$$E(v, h) = - \sum_{i=1}^{i^*} \sum_{j=1}^{j^*} W_{ij} v_i h_j - \sum_{i=1}^{i^*} a_i v_i - \sum_{j=1}^{j^*} b_j h_j \quad (8.4)$$

where

- Bias = $-(a^T v + b^T h)$
- a is the bias for the visible units
- b is the bias for the hidden units
- i^* is the number of visible units
- j^* is the number of hidden units.

The joint probability for visible and hidden units can be defined as

$$P(v, h) = \frac{1}{Z} e^{-E(v, h)} \quad (8.5)$$

where Z is the partition function defined as

$$Z = \sum_{v,h} e^{-E(v,h)} \quad (8.6)$$

Through marginalising the hidden vector h we can find the probability of the vector v as

$$P(v) = \sum_h P(v,h) = \frac{1}{Z} \sum_n \exp[-(E(v,h))] \quad (8.7)$$

As there is no connection in the hidden unit, the binary state h_j of hidden unit j is set to 1 with the probability

$$p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{i,j}) \quad (8.8)$$

Given a hidden vector v , we can easily calculate the step of visible units:

$$p(v_j = 1|h) = \sigma(b_j + \sum_i h_i w_{i,j}) \quad (8.9)$$

where $\sigma(x)$ is the sigmoid function. Using equations (8.8), (8.9) and Gibbs sampling techniques, we can easily update the visible unit vectors and hidden unit vectors. The weight function can also be improved by using the following equation:

$$\delta w_{i,j} = \epsilon(< v_i, h_j >_{\text{data}} - < v_i, h_j >_{\text{model}}) \quad (8.10)$$

Computing $< v_i, h_j >_{\text{data}}$ is comparatively easy, whereas the computation of the value $< v_i, h_j >_{\text{model}}$ is very difficult. The value of $< v_i, h_j >_{\text{model}}$ can be calculated by sampling methods like Gibbs, Contrastive Divergence (CD), Persistent Contrastive Divergence (PCD) and Free Energy in Persistent Contrastive Divergence (FEPCD).

We know that a Deep Belief Network (DBN) is constructed by stacking RBM models, acting as a skeleton for the construction of the DBN. In our model, we use 4 RBM layers, RBM-1, RBM-2, RBM-3 and RBM-4. RBM-1 has 18 inputs, because we have selected 18 features. Furthermore this RBM has 50 output units. Both RBM-2 and RBM-3 have 50 input units and 50 output units. Lastly RBM-4 has 50 input units and 2 output units, as we classify our data into two classes. The whole procedure is presented in Figure 8.5.

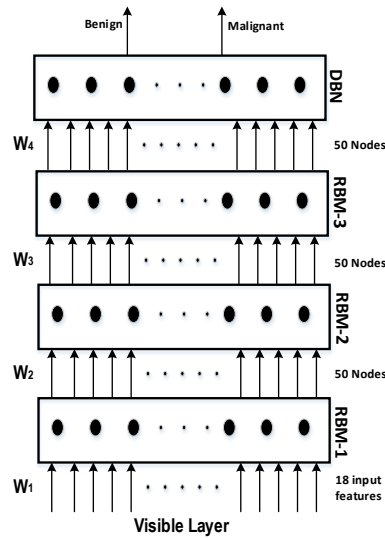


Figure 8.5: DBN model for analysis of the data

The input is first fed to the visible layer, which passes its input to the first RBM named RBM-1. The data moves back and forth between the RBM-1 layer and the visible layer until RBM-1 reaches some final decision. For updating the weight values and the neuron values, the network utilises equations (8.8), (8.9) and (8.10) for calculating the final values. As RBM-1 finally calculates its values, it passes these to the next hidden layer known as RBM-2. In this case, RBM-1 works as a visible layer for RBM-2.

Table 8.1: Detailed description of each block of the machine

Parameters	RBM-1	RBM-2	RBM-3	RBM-4	Output Layer
Input by Output	18 by 50	50 by 50	50 by 50	50 by 50	
No. of epochs	50	50	50	30	50 by 2
Sampling Method	CD	CD	CD	CD	

This same procedure is carried on throughout the network. As the network analysis proceeds, the weight value W_1 is developed between the visible layer and layer RBM-1. The weight value W_2 is developed between the RBM-1 layer and layer RBM-2. The weight

value W_3 is developed between the RBM-2 layer and RBM-3 layer, and the weight value W_4 is developed between the RBM-3 layer and the DBN layer. In our model we have used back propagation for fine tuning all the parameters along with the weight values, these being $W_1 + \epsilon_1$, $W_2 + \epsilon_2$, $W_3 + \epsilon_3$ and $W_4 + \epsilon_4$. All the particulars of our model and its sampling method are summarised in Table 8.1.

8.5 Contrast-Enhancement

The background image information of the histopathological images coexists with the foreground image information, and also the images suffer from poor contrast. To overcome these issues we have implemented the contrast-enhancement technique of [250] with modifications such as:

- Step-1: Background Subtraction

At first the original image information is subtracted from the non-uniform background information, separated using a low-pass Gaussian filter with standard deviation σ_1 . Depending on the value of σ_1 , Step-1 (shown in Figure 8.6) successfully removes the background variations globally.

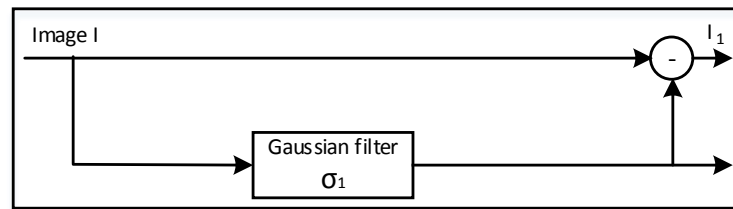


Figure 8.6: Block Diagram of Step-1.

- Step-2: Local adjustment

To improve the contrast information locally, the output image from Step-1 is divided pixel-wise by the variance of its spatial neighbour to minimise the contrast. Dividing

the whole image by the standard deviation σ_2 may amplify the noise inside the images, which degrades valuable image information. Step-2 shown in Figure 8.7.

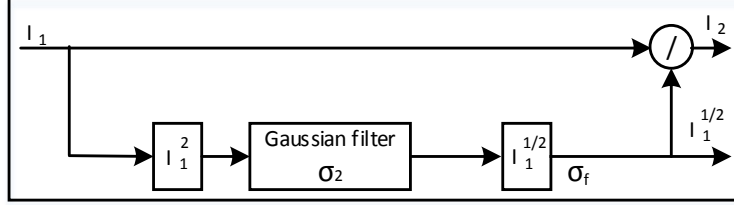


Figure 8.7: Block Diagram of Step-2.

- Step-3: Noise Control

To reduce the noise amplification Khan et al. [251] proposed a correction factor M:

$$M = 1 - \exp\left(-\frac{\sigma_f^p}{pC^p}\right) \quad (8.11)$$

where $p = 2$. Factor M multiplies the output of Phase-2. Here, σ_f is the local standard deviation and C is a user-defined parameter which controls the background noise. Step-3 has illustrated in Figure 8.8.

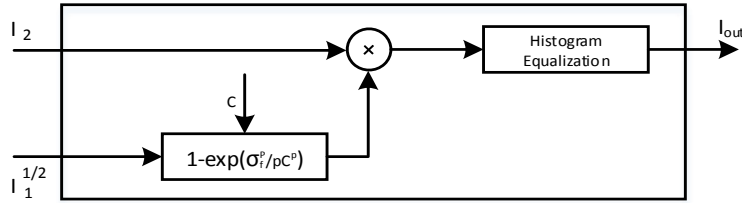


Figure 8.8: Block Diagram of Step-3.

The overall algorithm for the normalisation task (Algorithm-1) is shown below.

Algorithm 3 Proposed Contrast Enhancement Method

1. Input the image I .

2. Step-1:

- Calculate I_{σ_1} using Gaussian filter $\mathcal{J}_{\sigma_1}(\cdot)$: $I_{\sigma_1} \leftarrow \mathcal{J}_{\sigma_1}(I)$.
- Calculate I_1 where $I_1 = I - I_{\sigma_1}$.

3. Step-2:

- First find I_1^{root} from $I_1^{\text{root}} \leftarrow I_1^{\frac{1}{2}}$.
- Filter the image I_1^{root} n times with n different values of σ_2 with the Gaussian filter $\mathcal{J}_{\sigma_2}^{(n)}(\cdot)$. The values of all the available filtered images can be represented by the set $I_1^{\sigma_2(\text{all})} = \{I_1^{\sigma_2(1)}, I_1^{\sigma_2(2)}, I_1^{\sigma_2(3)}, \dots, I_1^{\sigma_2(n)}\}$. Here $I_1^{\sigma_2(n)}$ represents the filtered image I_1^{root} with the Gaussian filter $\mathcal{J}_{\sigma_2}^{(n)}(\cdot)$.
- $\sigma_2(\text{all}) = \{(\frac{\sigma_1}{k}) \text{ to } (\frac{\sigma_1}{1}) \text{ with increment } \Delta\}$; Here $\Delta \in \mathbb{R}_+$ and $k > l$ and $n = |\sigma_2^{\text{all}}|$.
- Select $\{I_1^{\sigma_2(\text{max})}\}$: $\{I_1^{\sigma_2(\text{max})}\} = \max\{I_1^{\sigma_2(\text{all})}\}$
- Select R_{max} : $R_{\text{max}} \leftarrow \max_{\text{pixel strength}}\{I_1^{\sigma_2(\text{max})}\}$
- Divide the image pixel-wise : $I_2 = \frac{I_1}{\{I_1^{\sigma_2(\text{max})}\}}$

4. Step-3:

- Calculate C : $C = R_{\text{max}} \times t$, where t is a user-defined value.
- Calculate value of M : $M = 1 - \exp(-\frac{\{I_1^{\sigma_2(\text{max})}p\}}{p \times C^p})$.

5. Calculate I_{norm} Image as: $I_{\text{norm}} = I_2 \times M$.

6. Perform histogram equalisation on the image I_{norm} and find out the I_{out} .

8.6 Feature-Extraction

One of the important steps of image classification is extracting the features from the images. Consider $f^{RGB}(u, v) = \{f^R(u, v), f^G(u, v), f^B(u, v)\}$ be an RGB image, here R, G, B represents the Red, Green, and Blue channel information. From the image $f^{RGB}(u, v)$, Tamura features vector T^R , T^G and T^B has been extracted from each of the respective channels shown as Figure 8.9.

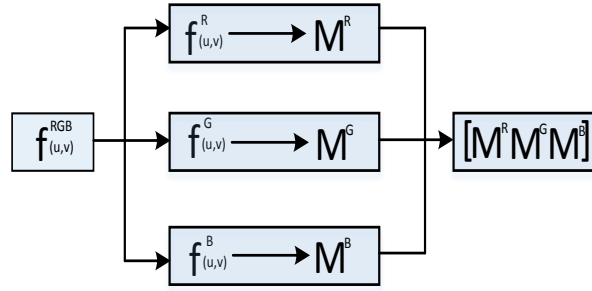


Figure 8.9: Tamura features extraction from the three different channels

1. Coarseness

Fineness of texture is measured by Coarseness. The measure of Coarseness is influenced by the scale as well as the duplication percentage of the components within that area. The largest size of the texture is also identified by Coarseness [252]. To calculate the coarseness within the image, average values are calculated at all the available points by varying the window size. Centred at the point (u, v) and for a window of size $2^k \times 2^k$, the average value can be formulated as

$$\mathcal{A}_k(u, v) = \sum_{i=u-2^{k-1}}^{u+2^{k-1}-1} \sum_{j=v-2^{k-1}}^{v+2^{k-1}-1} \frac{f(i, j)}{2^{2k}}. \quad (8.12)$$

$2^k \times 2^k$ non-overlapping neighbouring-window average variations have been calculated in both the horizontal and vertical directions:

$$\mathcal{C}_{k,h}(u, v) = |\mathcal{A}_k(u + 2^{k-1}, v) - \mathcal{A}_k(u - 2^{k-1}, v)| \quad (8.13)$$

$$\mathcal{E}_{k,t}(u, v) = |\mathcal{A}_k(u, v + 2^{k-1}) - \mathcal{A}_k(u, v - 2^{k-1})| \quad (8.14)$$

Irrespective of the direction, the value of k which maximises the output values is considered an optimal value. $\mathcal{S}_{\text{best}}$ is then calculated as

$$\mathcal{S}_{\text{best}} = 2^k \quad (8.15)$$

Finally the Coarseness \mathcal{C}_1 is calculated using

$$\mathcal{C}_1 = \frac{1}{m \times n} \sum_u^m \sum_v^n \mathcal{S}_{\text{best}}(u, v) \quad (8.16)$$

2. Contrast

The intensity within a texture contains a significant amount of information. Contrast represents the difference of the level of intensity within a texture. The following four factors are considered when Contrast is measured [9]:

- (a) The range of Gray level within an image
- (b) The Polarisation of the Gray-level distribution
- (c) Sharpness of edges
- (d) Period of repeating patterns.

Considering the above four factors Contrast can be defined as

$$\mathcal{C}_2 = \frac{\sigma}{(\alpha_4)^n} \quad (8.17)$$

where

$$\alpha_4 = \frac{\mu_4}{\sigma_4} : \text{is known as kurtosis}$$

μ_4 : Fourth moment about the mean

σ^4 : Variance²

3. Directionality

Directionality is a global property that refers to the shape of texture primitives and where they are placed within a specific region [9], [253], [254].

$$\mathcal{D}_1 = 1 - \text{r.n.o.} \sum_o^{\text{n.o.}} \sum_{\phi \in \text{w.o.}}^{\text{m}} (\phi - \phi_o)^2 \cdot \mathcal{H}_d(\phi) \quad (8.18)$$

\mathcal{H}_d : is the local direction histogram

n.o. : is the number of peaks of \mathcal{H}_d

w.o. : o is the range of the o^{th} peak between valleys

ϕ_o : o is the o^{th} peak position of \mathcal{H}_d

4. Line-likeness

Let $\mathcal{P}_{\text{Dd}}(\text{i,j})$ represent a directional co-occurrence matrix, where each element of this matrix is defined as "the relative frequency with which two neighbouring cells separated by a distance d along the edge direction occurs" [9], [253]. Co-occurrences in the same direction are weighted by $+1$, and co-occurrences with directions perpendicular to each other are weighted -1 . Using $\mathcal{P}_{\text{Dd}}(\text{i,j})$ the Line-likeness can be measured as [253]:

$$\mathcal{L} = \frac{\sum_i^n \sum_j^n \mathcal{P}_{\text{Dd}}(\text{i,j}) \cos(\text{i-j})(\frac{2 \times \pi}{\text{n}})}{\sum_i^n \sum_j^n \mathcal{P}_{\text{Dd}}(\text{i,j})} \quad (8.19)$$

5. Regularity

Regularity can be defined as

$$\mathcal{R}_1 = 1 - \text{r}(\sigma_{\text{crs}} + \sigma_{\text{con}} + \sigma_{\text{dir}} + \sigma_{\text{lin}}) \quad (8.20)$$

where r is the normalising parameter [9].

6. Roughness

According to the results of Tamura et al.'s experiments, a combination of coarseness

and contrast best aligns with the psychological results [9]:

$$\mathcal{R}_2 = \mathcal{C}_1 + \mathcal{C}_2 \quad (8.21)$$

8.7 Results and Discussion

We have utilised the BreakHis dataset for our experiments, where the dataset is grouped into $m = \{40\times, 100\times, 200\times, 400\times\}$ groups where \times represents the magnification factor. Each of the images in this dataset is RGB in nature and 700×460 pixels in size. We have used Tamura features as attributes and extracted the features from all the channels, which produces a total of 18 features.

The experiments have been performed on each of the individual groups of the dataset separately; 70 percent, 15 percent, and 15 percent of the data have been used for the training, validation and testing purposes, respectively. Let each group in the dataset be represented by the set X^m .

$$X^m = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \dots \\ \mathbf{x}_{T_m} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_{T_m} \end{bmatrix} = \begin{bmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & \dots & x_1^S \\ x_2^1 & x_2^2 & x_2^3 & \dots & \dots & x_2^S \\ x_3^1 & x_3^2 & x_3^3 & \dots & \dots & x_3^S \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{T_m}^1 & x_{T_m}^2 & x_{T_m}^3 & \dots & \dots & x_{T_m}^S \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_{T_m} \end{bmatrix}$$

Here the value of S is equal to 18. T_m represents the total data of the group

$$T_m = tr_m + ts_m + tv_m \quad (8.22)$$

tr_m = training data of group m ;

ts_m = test data of group m ;

tv_m = validation data of group m .

$y_i \in \{\text{Benign}, \text{Malignant}\}$.

The results of all the experiments of this chapter have been evaluated through the Confusion Matrix (CM) and a few other performance-measuring parameters. A two-dimensional table which illustrates the performance of a classifier is known as a CM [43]. If a classifier provides 100% accuracy performance then all the non-diagonal elements of the CM will be zero [255]. Table 8.2 shows a graphical representation of a CM for a binary classifier along with a few performance-measuring parameters.

- $TPR = TP / (TP + FN)$
- $TNR = TN / (TN + FP)$
- $Accuracy = (TP + TN) / (TP + FN + TN + FP)$

		Hypothesized Class	
True Class	Benign	True Positive (TP)	False Negative (FN)
	Malignant	False Positive (FP)	True Negative (TN)
		Benign	Malignant

Table 8.2: Few Performance measuring parameters along with CM

8.7.1 Results and Comparison

In Figures 8.10 and 8.11, the a, b, c and d images show the Train, Validation, Test and Overall performance when we use the $40\times$, $100\times$, $200\times$ and $400\times$ datasets for Algorithm-1 and Algorithm-2, respectively. When we use the $40\times$ database and Algorithm-2, the Train, Test, Validation and overall accuracies remain almost the same, at around 88.7%. When we use the $100\times$ dataset, the Test shows less accuracy than the Train and Validation performance. When we use the $200\times$ dataset, the Train, Validation, Test and Overall accuracies are 89.4%, 86.3%, 87.7% and 86.8%, respectively. When we use the $400\times$ database, the overall accuracy achieved is around 88.4%.

These Confusion Matrices also show that, when we utilised Algorithm-1 for the $40\times$,

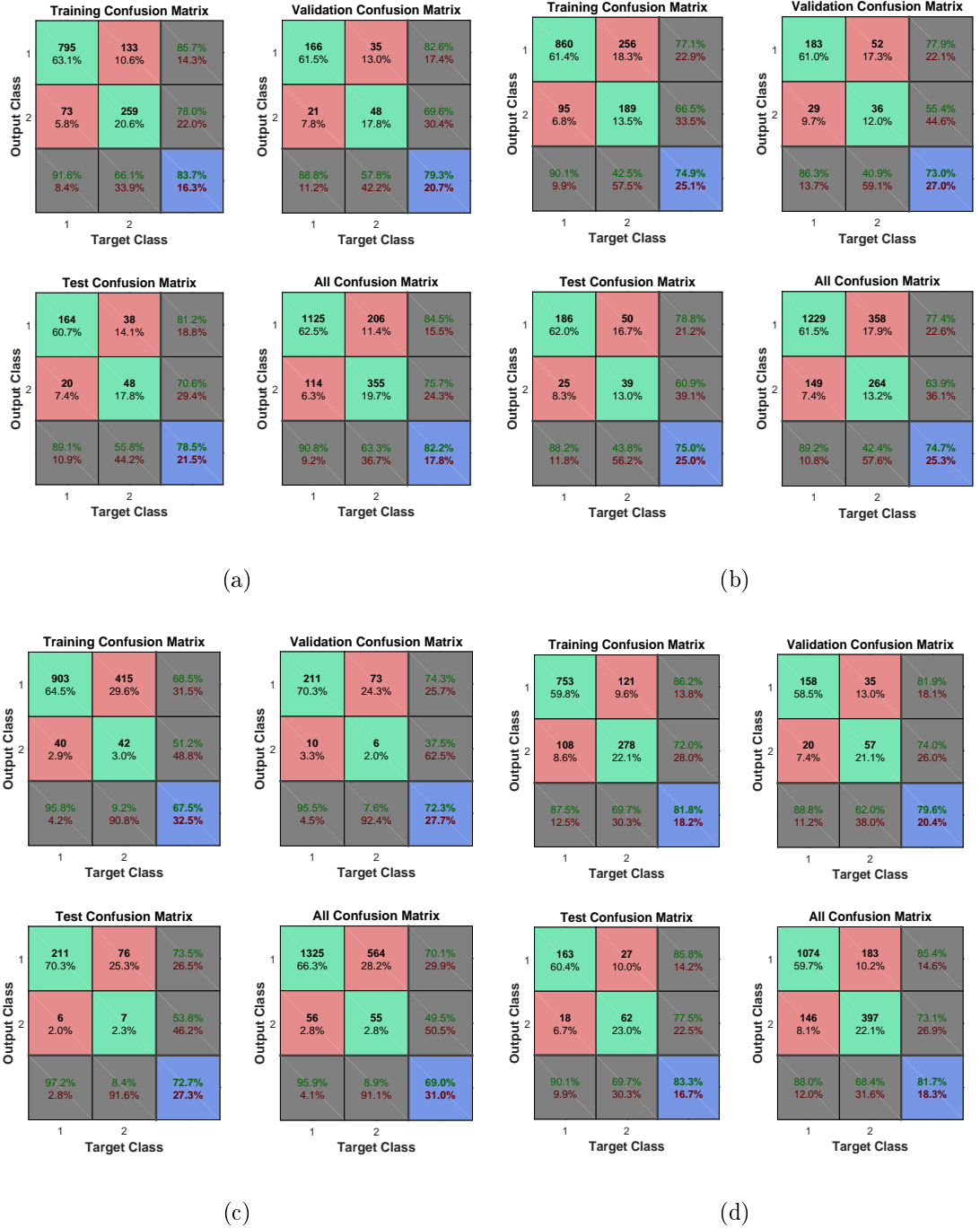


Figure 8.10: (a), (b), (c) and (d) represent the Confusion Matrices for Algorithm-1 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.

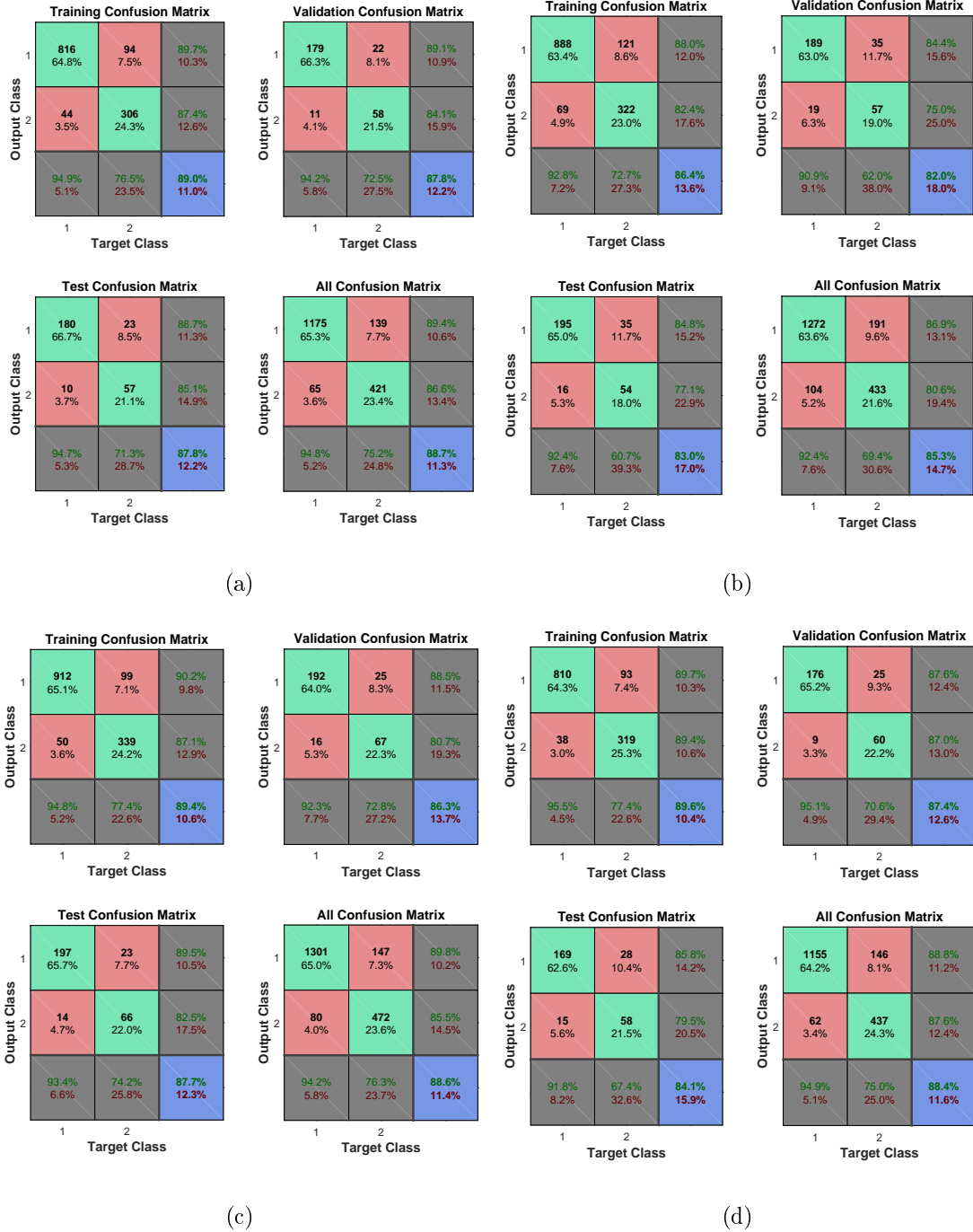


Figure 8.11: (a), (b), (c) and (d) represent the Confusion Matrices for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.

100 \times , 200 \times and 400 \times magnification-factor database, 15.5%, 22.60%, 29.90% and 14.6% of the Malignant images have been misclassified as Benign images. However, 24.30%, 36.10%, 50.50% and 26.9% of the Benign images have been wrongly classified as Malignant images. The overall accuracy achieved for the 40 \times , 100 \times , 200 \times and 400 \times cases was 82.20%, 74.70%, 69.00% and 81.70%, respectively; for all magnification factors a greater percentage of the database has been misclassified as Benign. When we utilised Algorithm-2, 10.60%, 13.10%, 10.02% and 11.20% of the Malignant data were misclassified as Benign images for the 40 \times , 100 \times , 200 \times and 400 \times cases, respectively. On the other hand, 13.40%, 19.40%, 14.50% and 12.40% of the data has been classified as Malignant though they are originally Benign images for the 40 \times , 100 \times , 200 \times and 400 \times cases, respectively.

Performance

The Mean-Square Error (MSE) assesses the quality of a model and a good classifier is expected to have a small MSE. Let θ be the predicted value, θ' be the observed value for n observations, then the MSE error can be defined as

$$\text{MSE} = \frac{1}{n} \left[\sum_{i=1}^n (\theta_i - \theta'_i) \right]. \quad (8.23)$$

Figure 8.12, a, b, c, and d illustrate the performance of the 40 \times , 100 \times , 200 \times and 400 \times datasets when we use Algorithm-1. Figure 8.13, a, b, c, and d depict the performance of the 40 \times , 100 \times , 200 \times and 400 \times datasets when we use Algorithm-2. Table 8.3 summarises the MSE values and the required number of epochs to achieve that value.

Table 8.3 and Figures 8.12 and 8.13 show that, for Algorithm-1, the best MSE values are achieved when we use the 400 \times magnification factor, and it takes 209 epochs. However, when we have recourse to Algorithm-1 and the 200 \times magnification factor dataset the model requires 26 epochs to achieve an MSE of 0.18074. Though it requires fewer epochs, it performs worse than all the other datasets when we exploit Algorithm-1. When we

Table 8.3: MSE values and the corresponding epoch values

Algorithm	Magnification Factor	MSE	Epoch
Algorithm-1	40×	0.14481	209
	100×	0.16528	134
	200×	0.18074	26
	400×	0.13813	238
Algorithm-2	40×	0.09941	373
	100×	0.09384	236
	200×	0.09384	254
	400×	0.09948	483

utilise Algorithm-2 almost all the datasets show the same kind of MSE, which lies in between 0.09384 and 0.09948. However, when we implement the 400× database it requires 483 epochs, which is larger than for the other three datasets.

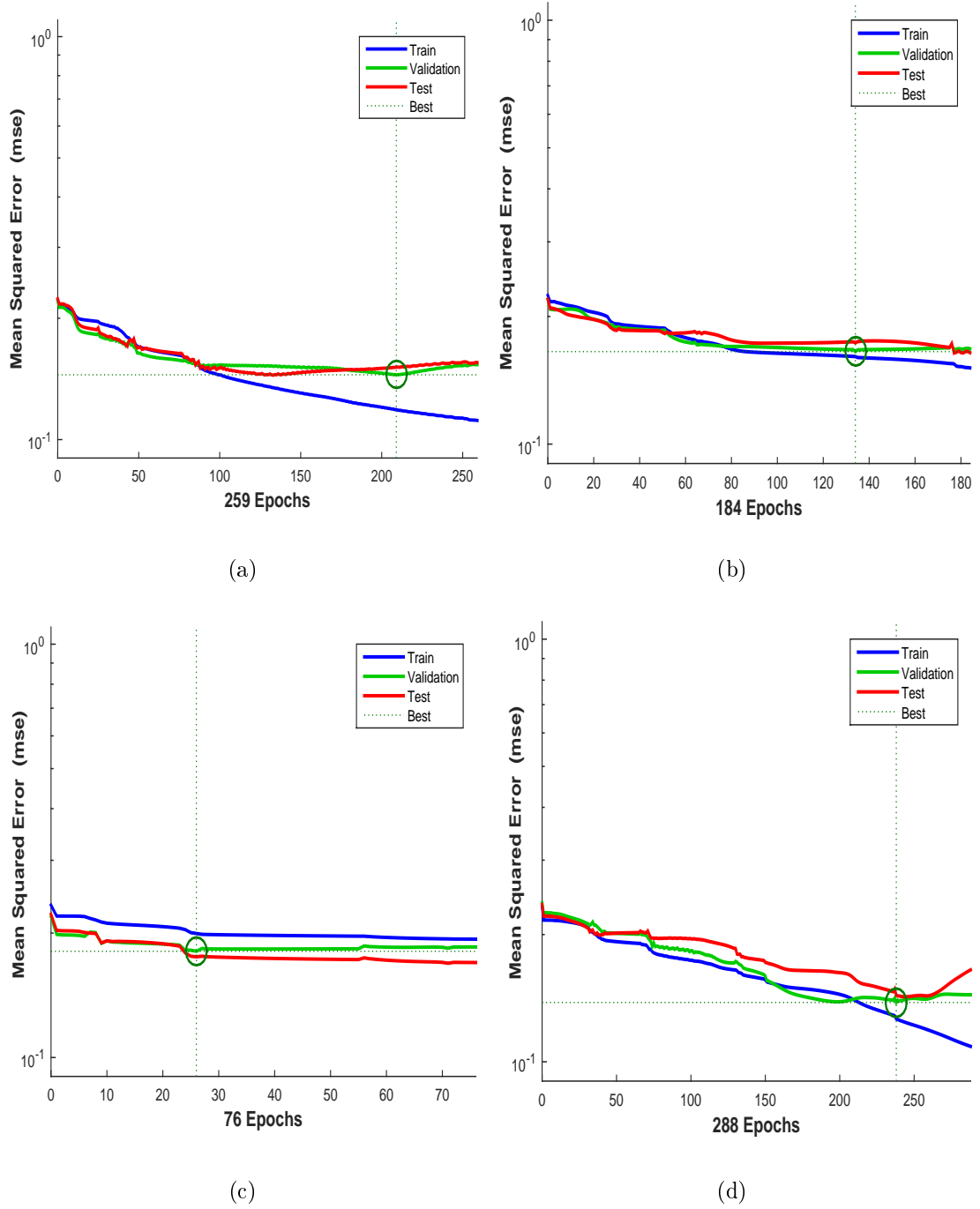
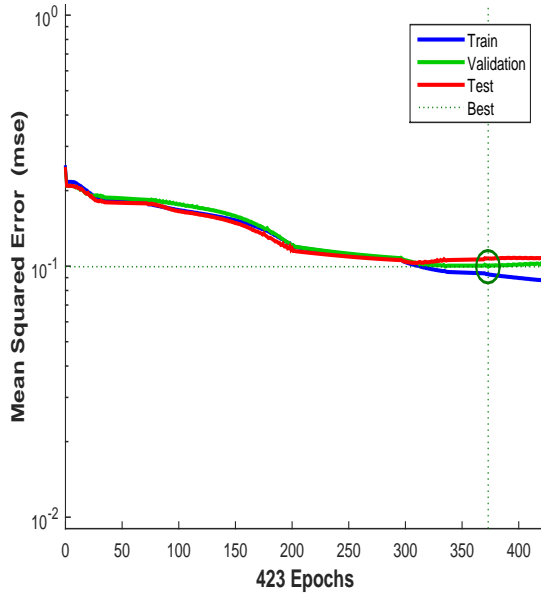
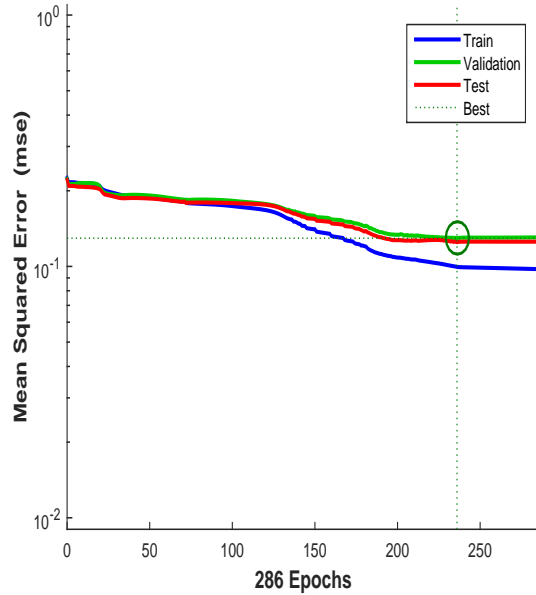


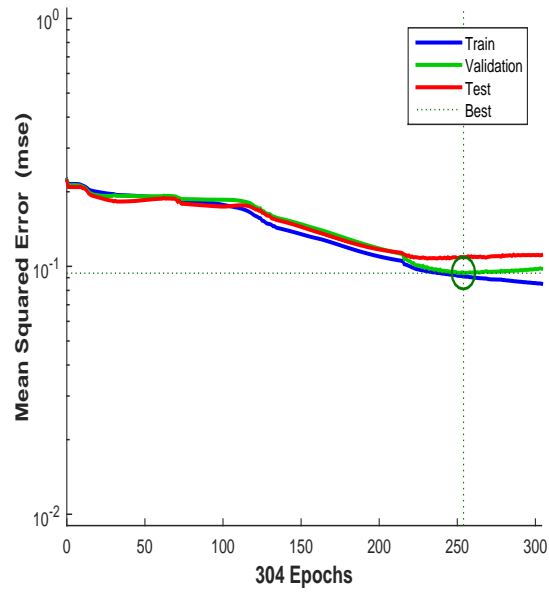
Figure 8.12: (a), (b), (c) and (d) represent the performance analysis for Algorithm-1 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.



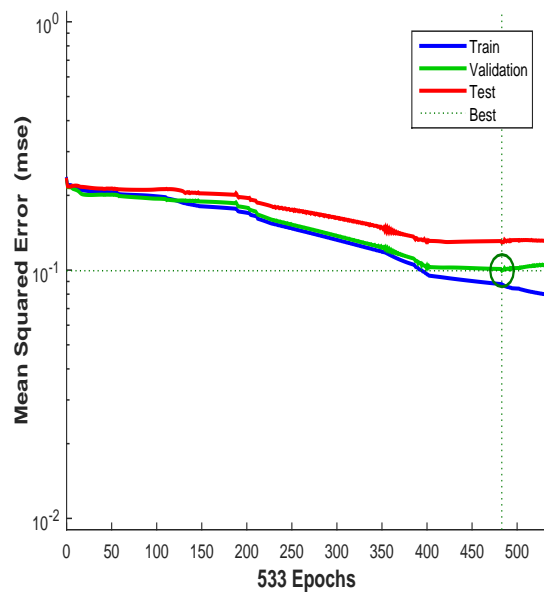
(a)



(b)



(c)



(d)

Figure 8.13: (a), (b), (c) and (d) represent the performance analysis for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.

ROC curves

ROC curves show the False Positive Rate and True Positive Rate performance. The best performance is achieved at the top-most left position. That position indicates that the False Positive Rate is 0 and the True Positive Rate is 1, which also indicates that the True Negative rate is 100.00%. Figures 8.14 and 8.15 show ROC curves for the Algorithm-1 and the Algorithm-2, respectively.

So far, very little work has been done on classifying the BreakHis dataset. Fabio A. Shanol et al. used the Local Binary Pattern (LBP), Local Plane Quantization (LPQ), Gray-Level-Co-occurrence Matrix (GLCM), Parameter Free Threshold Adjacency Statistics (PFTAS) methods for Feature-Extraction. These authors applied four different classifiers: 1-Nearest Neighbor (1-NN), Quadratic Linear Analysis (QDA), Support Vector Machine (SVM) and Random Forest (RF). Overall they achieved the best performance when they used the PFTAS descriptor and SVM classifier, and their achieved performance Accuracy was $85.1 \pm 3.1\%$. As a descriptor, we use Tamura features. Our proposed Algorithm-1 and Algorithm-2 both use the RBM method for image classification. When we use Algorithm-1, the overall accuracy achieved is 82.20%, 74.70%, 69.00% and 81.70% for the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively, while Algorithm-2 gave 88.70%, 85.30%, 88.60% and 88.40% accuracy for the $40\times$, $100\times$, $200\times$ and $400\times$ databases, respectively.

In [205] the performance has been evaluated through the accuracy measure. However, in this chapter we have found the ROC information and the error performances with the epoch.

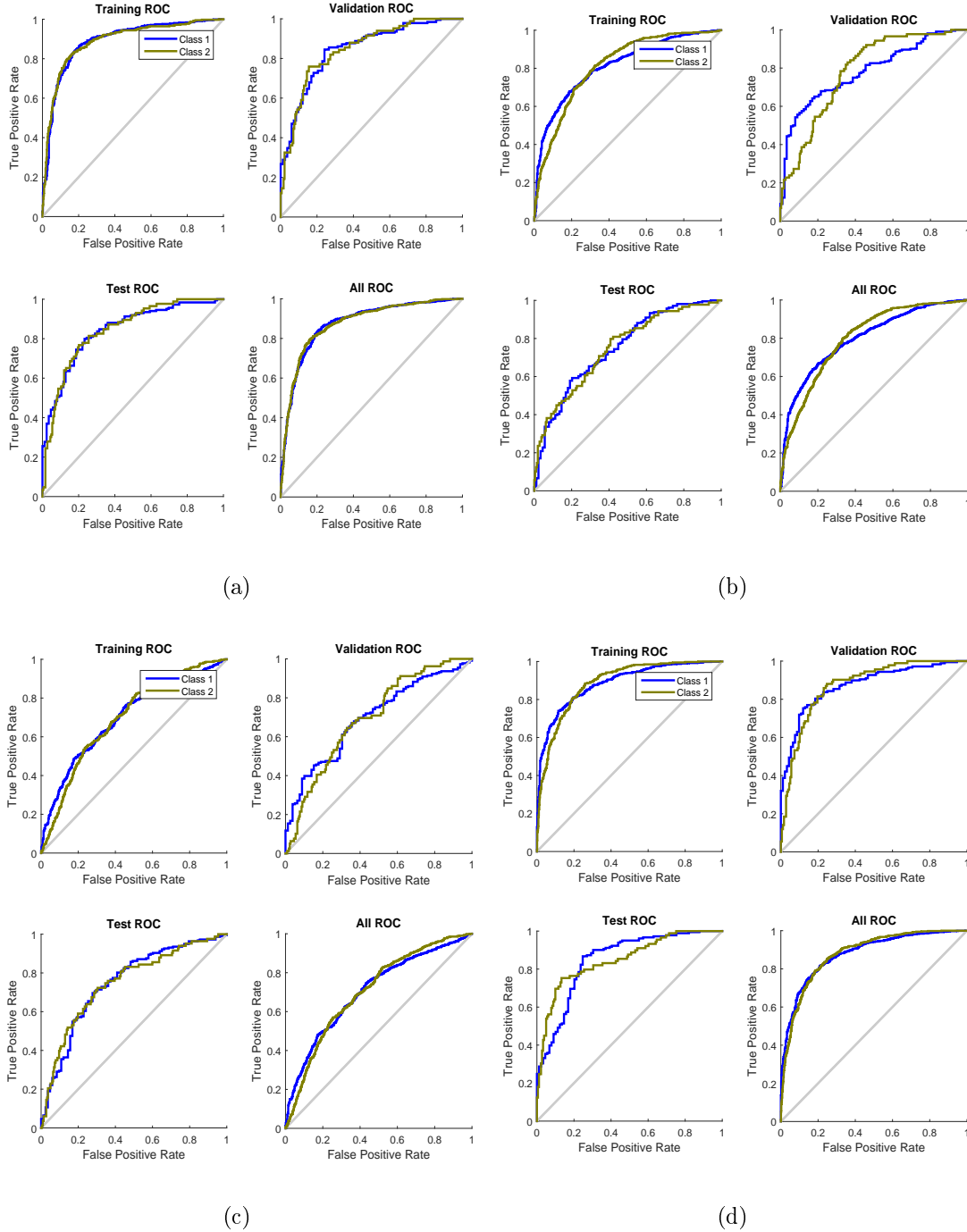


Figure 8.14: (a), (b), (c) and (d) represent the ROC curves for Algorithm-1 when we utilise the 40 \times , 100 \times , 200 \times and 400 \times datasets, respectively.

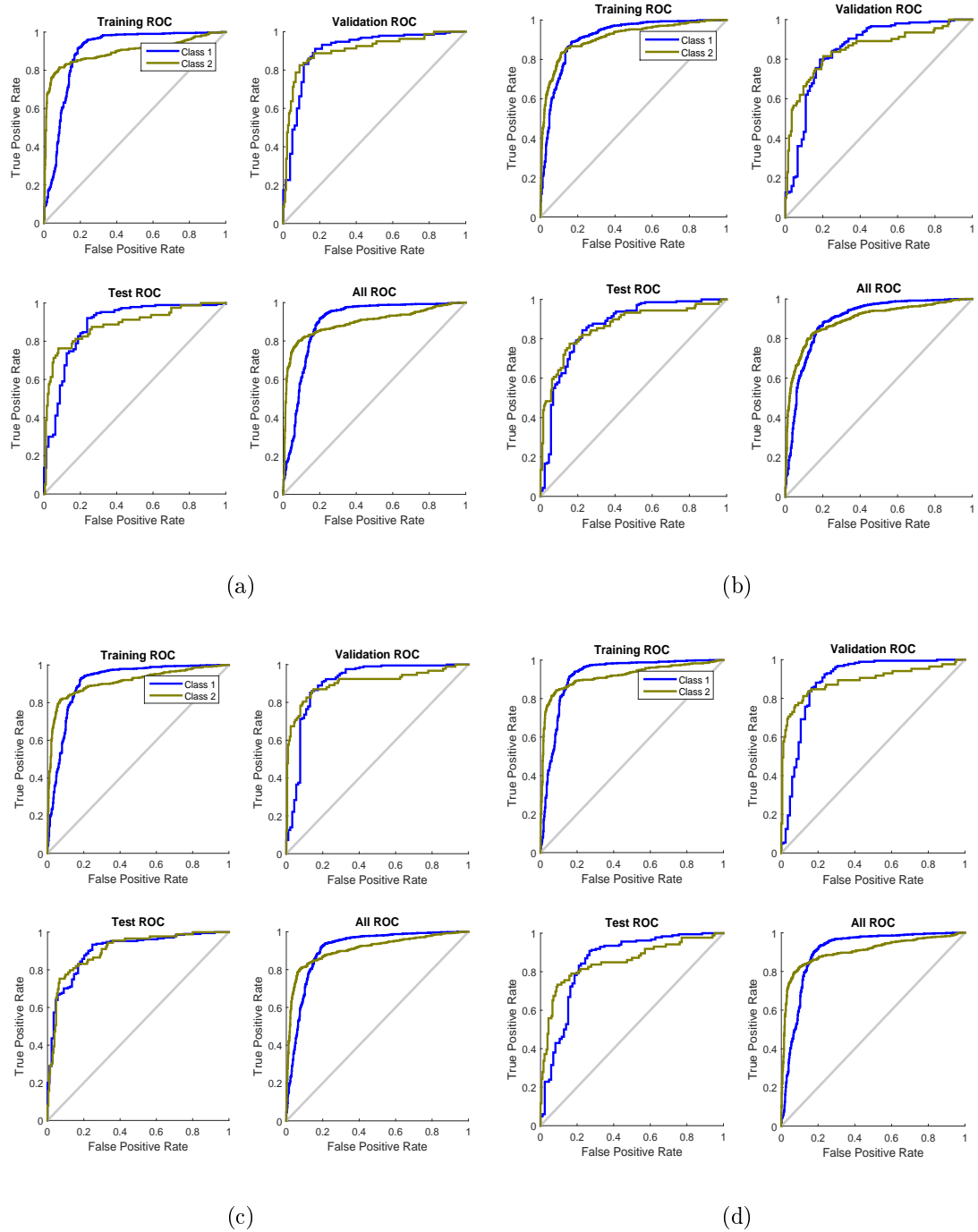


Figure 8.15: (a), (b), (c) and (d) represent the ROC curves for Algorithm-2 when we utilise the $40\times$, $100\times$, $200\times$ and $400\times$ datasets, respectively.

Table 8.4: Comparison of results using our proposed algorithm and other algorithms

Descriptor	Classifier	Magnification Factor and Accuracy %			
		40×	100×	200×	400×
CLBP [205]	SVM	77.4±3.8	76.4±4.5	70.2±3.6	72.8±4.9
GLCM [205]	RF	73.6±1.5	76.0±1.9	82.4±2.3	79.8±2.5
LBP [205]	SVM	74.2±5.0	73.2±3.5	71.3±4.0	73.1±5.7
LPQ [205]	1-NN	72.8±4.9	71.1±6.4	74.3±6.3	71.4±5.2
ORB [205]	QDA	74.4±1.7	66.5±3.2	63.5±2.7	63.5±2.2
PFTAS [205]	SVM	81.6±3.0	79.9±5.4	85.1±3.1	82.3±3.8
Algorithm-1	RBM	82.2	74.7	69.0	81.7
Algorithm-2	RBM	88.7	85.3	88.6	88.4

8.8 Conclusion

In this chapter we have proposed an automatic BC image classifier framework which has been constructed using state-of-the art Deep Neural Network techniques. Instead of using raw images we have utilised Tamura features, as they provide textural information. As a deep-learning tool we have implemented an unsupervised Restricted Boltzmann Machine which contains four layers and is guided by a supervised backpropagation technique. For the back-propagation, Scaled Conjugate Gradient techniques have been utilised. We have performed our experiments on the BreakHis dataset and obtained 88.7%, 85.3%, 88.6% and 88.4% Accuracy for the dataset of 40×, 100×, 200× and 400× magnification factors, respectively. Most of the experiments on the BreakHis dataset judged the performance on the basis of Accuracy, however, in this chapter we have also considered TPR, FPR values along with a detailed description of the ROC curves. The error performance as a function of the epoch is also explained in detail. This chapter shows that the RBN method is very effective for automatic breast-cancer image diagnosis. However, in the future the combination of CNN and RBM will enhance the classification performance.

Chapter 9

Histopathological Breast-Cancer Image Classification with Feature Prioritisation

9.1 Abstract

Breast-Cancer image classification has always been a challenging task. Among all the BC images, Histopathological images always provide valuable information about the cancer. This kind of image analysis requires a specialist opinion, but sometimes specialists disagree about the final outcome. Apart from this, this kind of image analysis is very time-consuming and needs great patience. However, the modern Computer Aided Diagnosis (CAD) system can help the doctors and the physician to justify their decisions, which gives the patient more satisfaction. To date various image-classification algorithms are available for BC image classification, but in this chapter we have mainly utilised the "Ex-

In review as: A. A. Nahid, A. Mikaelian, M. A. Mehrabi and Y. Kong, "Histopathological Breast-Cancer Image Classification with Feature Prioritisation", *BioMed Research International*, Hindawi, pp. —, —.

treme Gradient Boosting" algorithm (XGBoost). It is the most-recent advanced version of Gradient Boosting, for the classification of a set of Histopathological breast images into Benign and Malignant images. As a feature we have utilised Tamura, Histogram, Local Binary Pattern (LBP) and Haralick features for the BC image classification. The mathematical structure of the XGBoost algorithm contains a few parameters which require fine-tuning for better performance of the classifier model. In this chapter, after adjusting the parameters such as the depth of tree, number of trees, and the learning rate, this chapter achieved 98.22%, 97.90%, 98.14%, 98.06% and 98.64% Accuracy, Recall, Precision, F-measure, and Specificity, respectively when Tamura features are utilised.

9.2 Introduction

Unwanted growth of cells in a body damages the natural working mechanism, which leads to cancer. BC is of particular concern to women as, due to the anatomical structure of the female body, they are more vulnerable to BC than men. As an example, Figure 9.1 shows the statistics concerning BC deaths in Australia for the last decade. These statistics show that the female death rate is far greater than the male death rate due to BC. It is an example that we may consider as representing the current situation throughout the world.

Lobules, ducts, nipples and fatty tissue are the main structural components of the female breast. Normally BC starts from the lobules and ducts, then the created cancer infects the whole body. Cancer can be divided into a number of classes, however, in a broad sense cancer can be classified into two types:

- Benign: which is not life-threatening
- Malignant: which can be life-threatening in the future.

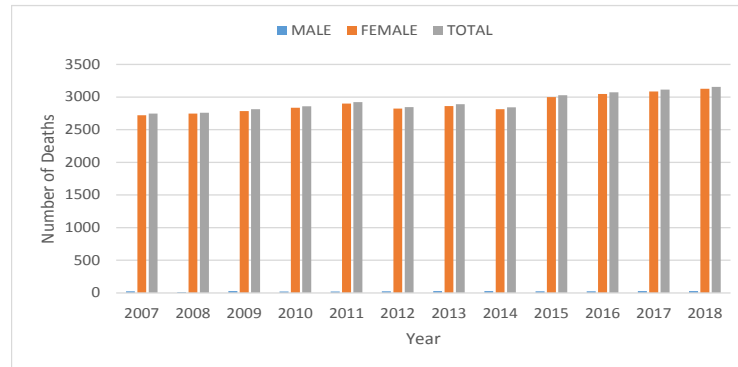


Figure 9.1: Death statistics for Breast Cancer over the last decade in Australia

Detection of cancer is the prerequisite state for cancer treatment. Both early and accurate detection of cancer can save many people's lives, or at least alleviate their miserable condition. Along with other techniques, the investigation of cancer largely depends on biomedical photographic techniques such as Histopathological images. Proper investigation of these types of images to distinguish between the Benign and the Malignant images depends on some prerequisite pre-processing steps, as well as the quality of the images and the expertise of investigators. With cancer, investigation of the images is always a time-consuming and challenging task. However, with the advances made in modern digital photographic techniques the medical community has become largely reliant on these images for cancer investigation and classification. Different methods, techniques and steps have been used for BC image classification, which follows some predefined procedure such as Feature-Selection along with the appropriate model utilisation.

The Gray-Level Co-occurrence Matrix (GLCM) along with the Random Forest (RF) method has been employed for BC image classification by different investigators. An-gayarkanni et al. [118] and Beura et al. [256] have implemented the GLCM method for mammographic breast-image classification and obtained 99.50% and 98.0% Accuracy, respectively. Bruno et al. [120] utilised the Curvelet Transform (CT) and LBP together for Histopathological image classification using the RF algorithm. They also accessed

the Analysis of Variance (ANOVA) method for the Feature-Selection and achieved an Accuracy in the range of 91.00% to 100.00%. Paul et al. [115] utilised Haralick Texture features along with the RF algorithm for Histopathological image classification and they obtained Recall and Precision values of 81.13% and 83.50%, respectively.

Rajakeerthana et al. [50] utilised GLCM, GLD, SRDM NGLCM, and GLRM Feature-Extraction techniques along with the Neural Network (NN) method for mammographic breast-image classification and obtained a 99.20% classification Accuracy. Atlas et al. [55] employed morphological features along with the NN method for mammographic breast-image classification and obtained around 97.50% Accuracy. Peng et al. [57] used Harlick and Tamura features with NN for mammographic image classification, and they also accessed feature reduction as suggested by Rough-Set theory. Variance contrast as well as auto-correlation of wavelet coefficients and the NN method for ultrasound breast-image classification were used by Chen et al. [62]; they obtained a Receiver Operating Characteristic (ROC) curve value of 0.9396. Jalalian et al. [58] utilised the GLCM method for mammogram image classification using the NN method and obtained classifier Accuracy, Sensitivity and Specificity of 95.20%, 92.40% and 98.00%, respectively.

Chang et al. [133] relied on textural features such as auto-correlation coefficients along with a Support Vector Machine (SVM) for breast-image classification. Their obtained Accuracy, Sensitivity and Specificity were 85.60%, 95.45%, and 77.86%, respectively. Kavitha et al. [143] utilised histogram information for mammogram image classification along with the SVM method; when using SVM with the linear kernel the obtained Accuracy, Sensitivity and Specificity were 98.00%, 100.00% and 96.00%, respectively, and when using weighted-feature SVM the obtained Accuracy, Sensitivity and Specificity were 90.00%, 100.00% and 75.00%, respectively. Kancham et al. implemented Tamura features, shape-based features and moment-invariant features to detect the abnormality of breast-images on the Mini-MIAS database. When they utilised Tamura features and the

SVM method together their obtained Accuracy was 78.46% [257]. Zhang et al. utilised fractional-fourier-transform information as features with an SVM along with Principal Component Analysis (PCA) for breast-image classification and the achieved Accuracy, Sensitivity and Specificity were $92.16 \pm 3.60\%$, $92.10 \pm 2.75\%$ and $92.22 \pm 4.16\%$, respectively [137]. Shirazi et al. [138] used GLCM features with ultrasound images. The Regions of Interest (ROI) were extracted to reduce the redundant complexity, while SVM and the mixed gravitational search algorithm (MGSA) served as classifiers for the image classification. Dheba et al. [140] utilised the Laws texture features for mammogram image classification and obtained 86.10% Accuracy. Zhou utilised GLCM and the Tamura technique to classify the DDSM data set and achieved a 69.00% Accuracy [258]. LBP features were utilised by Karl et al. for breast-image classification using an SVM classifier. They did their experiments on both the MIAS and DDSM databases [259].

The literature shows that a few studies have been performed on breast-image classification using features like LBP, GLCM, Histogram along with some conventional image-classifier models like SVM, NN. Most of the work has been performed on very well-known datasets like MIAS and DDSM along with some Histopathological image-based datasets. BreakHis is a very recent Histopathological breast-image dataset [7] and only a few image-classification studies have been performed based on this dataset. For the very first time, Spanhol et al. [7] undertook breast-image classification on this dataset where they utilised the LBP, Local Plane Quantization (LPQ), GLCM, Parameter-Free Threshold Adjacency Statistics (PFTAS) methods for the Feature-Extraction. They utilised four different classifiers: 1-Nearest Neighbor (1-NN), Quadratic Linear Analysis (QDA), SVM and RF. Overall they achieved the best performance when they used the PFTAS descriptor and SVM classifier, and their achieved accuracy was $85.1 \pm 3.1\%$.

Extreme Gradient Boosting (XGBoost) is a very recent advanced model for classification. To the best of our knowledge only a very small amount of work has been done

using it for BC image classification. In the case of the Histopathological-image (BreakHis) dataset, so far no work has been done on classification using the XGBoost algorithm. In this chapter we have classified the BreakHis dataset using the XGBoost algorithm, while we utilised Tamura, Harlick, Histogram and LBP features for the classification. We have also adapted a few Feature-Selection algorithms to find the best features for the image classification. We have organised our chapter as follows: Section 9.2 gives a brief introduction to BC image classification issues, Section 9.3 describes the classification methodology along with a detailed description of the XGBoost algorithm and some Feature-Selection techniques. Section 9.4 compares the classification results with the XGBoost algorithm and other algorithms, Section 9.5 explains and utilises the Feature-Selection methodology with the XGBoost algorithm to find the suitable feature sets. We conclude the chapter in Section 9.6.

9.3 Classification Methodology

Histopathological Benign and Malignant breast-image classification is always a challenging task due to the complexity of the image. Figures 9.2 a and b show Benign and Malignant images selected from the BreakHis breast-image dataset. In this chapter we have utilised supervised image-classification techniques for the BreakHis image dataset classification.

A conventional supervised BC image-classification procedure follows a set of predefined steps such as

- Selection of the BC image dataset.
- Feature-Extraction and Selection.
- Classifier Model.
- Classifier Output.

which is illustrated in Figure 9.3. Consider the dataset

$$\mathcal{D} = \{X_{i,j}, Y_i\} \quad (9.1)$$

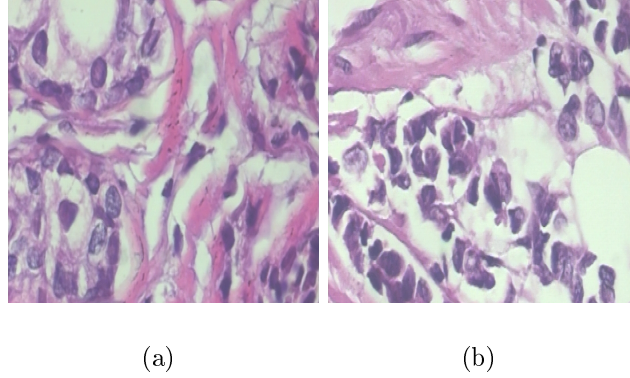


Figure 9.2: (a), (b) show Benign and Malignant images selected from the BreakHis dataset.

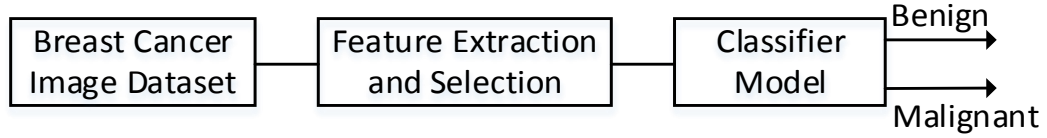


Figure 9.3: A very basic image-classification model

where $|\mathcal{D}| = n$ represents the total number of instances of the set. $X_{i,j}$ represents the feature vector of an instance with the corresponding labels $Y_i \in \{\text{Benign}, \text{Malignant}\}$; here $i = \{1, \dots, n\}$, $j = \{1, \dots, p\}$ and p represents the number of features. The total dataset has been splitted into a train and a test dataset. The model has been trained using the train dataset along with the corresponding labels:

$$\mathcal{F}'_x = \mathcal{F}_x(X_{i,j}, Y_i). \quad (9.2)$$

The test dataset been evaluated on the model \mathcal{F}'_x . Along with other classifier models we have utilised the most recent state-of-the-art XGBoost algorithm for image classification.

9.3.1 Extreme Gradient Boosting

The XGBoost algorithm follows the principle of the Gradient Boosting algorithm. Boosting means that a collection of weak learners can combine to produce a strong learner. Let

\mathcal{F}_x^t be the overall tree model after round $t - 1$ which can be defined as

$$\mathcal{F}_x^t = \begin{cases} 0, & \text{if } t = 0 \\ \mathcal{F}_x^{t-1} + h_x, & \text{if } t > 0 \end{cases} \quad (9.3)$$

Here h_x is the new tree which is added to \mathcal{F}_x^{t-1} to predict \mathcal{F}_x^t at time t with the consideration of minimising the objective function:

$$\mathcal{O}(\mathcal{F}_x^t) = \mathcal{L}(\mathcal{F}_x^t, \mathcal{F}_x^{t-1}) + \mathcal{C}(\mathcal{F}_x^t). \quad (9.4)$$

Here $\mathcal{C}(\mathcal{F}_x^t) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ where T represents the number of leaves, w represents the weight vector, γ is the tree-size penalty parameter and λ is the leaf-weight penalty parameter.

9.3.2 Feature-Extraction

An important stage of image-classification is extracting the features from the images. In the conventional image classification task, features are crafted locally using some specific rules and criteria. Four different Feature-Extraction techniques we have utilised are:

- Tamura
- Histogram
- Harlick Features
- Local Binary Pattern

Let $f^{\text{RGB}}(u, v)$ represent an RGB image, which contains 3 channels $f^{\text{R}}(u, v)$, $f^{\text{G}}(u, v)$, $f^{\text{B}}(u, v)$; here R, G and B represent Red, Green and Blue, respectively. For the Feature-Extraction we have employed all three channels, as shown in Figure 9.4. M^{R} , M^{G} and M^{B} represent the extracted feature vectors from the R, G and B channels, respectively.

Tamura

Inspired by psychological studies of human visual perception, Tamura features extract six different textural features, proposed by Tamura et al. [9]:

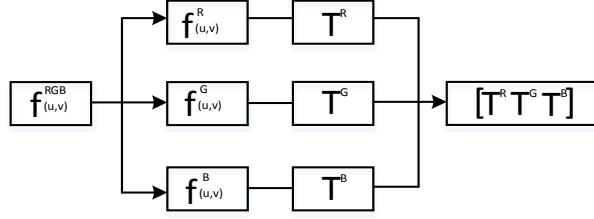


Figure 9.4: Overall Feature-Extraction from the three different channels

- **Coarseness:** Coarseness measure the fineness of a texture and also identifies the largest size of the texture [252]. Varying the kernel size, average values are calculated at each point of the image. Considering the target point as (g, l) , then the average value of that target point is calculated by varying the kernel size:

$$\mathcal{X}_k(g, l) = \sum_{i=g-2^{k-1}}^{g+2^{k-1}-1} \sum_{j=l-2^{k-1}}^{l+2^{k-1}-1} \frac{f(i, j)}{2^{2k}} \quad (9.5)$$

The average variation due to using a $2^k \times 2^k$ non-overlapping kernel window is calculated in both horizontal and vertical directions. Then, the value of k which maximises the output values is considered the optimal value \mathcal{S}_{opt} , and finally using this optimal value \mathcal{S}_{opt} the Coarseness \mathcal{C}_1 is calculated as

$$\mathcal{C}_1 = \frac{1}{m \times n} \sum_g^m \sum_l^n \mathcal{S}_{\text{opt}}(g, l) \quad (9.6)$$

- **Contrast:** The level of difference of intensity within a texture is defined as Contrast, can be defined as:

$$\mathcal{C}_2 = \frac{\sigma}{(\beta_4)^n} \quad (9.7)$$

where

$\beta_4 = \frac{\mu_4}{\sigma_4}$: is known as kurtosis

μ_4 : Fourth moment about the mean can be defined as $E[f^4(i, j)]$

σ^4 : Gray-Level distribution's standard deviation is represented as σ

- Directionality: Directionality refers to the shape of the texture primitives and their placing regulation within a specific region [9], [253], [254].

$$\mathcal{D}_1 = 1 - r \cdot n_o \cdot \sum_o^{n_o} \sum_{\phi \in w_o}^m (\phi - \phi_o)^2 \cdot \mathcal{T}_d(\phi) \quad (9.8)$$

\mathcal{T}_d : Histogram's local direction

n_o : is the number of peaks of \mathcal{T}_d

w_o : is the range of the o^{th} peak between valleys

ϕ_o : is the o^{th} peak position of \mathcal{T}_d

- Line-likeness: The element of a texture that is composed of lines can be described as Line-likeness. To calculate Line-likeness, a direction co-occurrence matrix consisting of elements $\mathcal{Q}_{\text{Dd}}(i,j)$, is generated [9], [253]. The following equation computes a measure of Line-likeness [253]:

$$\mathcal{L} = \frac{\sum_i^n \sum_j^n \mathcal{Q}_{\text{Dd}}(i,j) \sin(i-j)(\frac{2 \times \pi}{n} - \frac{\pi}{2})}{\sum_i^n \sum_j^n \mathcal{Q}_{\text{Dd}}(i,j)} \quad (9.9)$$

- Regularity: Regularity can be defined as

$$\mathcal{R}_1 = 1 - r(\sigma_{\text{crs}} + \sigma_{\text{con}} + \sigma_{\text{dir}} + \sigma_{\text{lin}}) \quad (9.10)$$

where r is a normalising parameter [9].

- Roughness: A combination of Coarseness and Contrast is represented as Roughness [9]:

$$\mathcal{R}_2 = \mathcal{C}_1 + \mathcal{C}_2 \quad (9.11)$$

Histogram

Let $f(u, v)^{\text{RGB}}$ represent an RGB image where R, G, and B stand for the Red, Green and Blue channel light information, respectively. Let $H(u, v)_L^{\text{R}}$, $H(u, v)_L^{\text{G}}$, $H(u, v)_L^{\text{B}}$ represent

the histograms of $f^R(u, v)$, $f^G(u, v)$, $f^B(u, v)$; here L represents the level of colour information where $L = \{2^0 \text{ to } 2^8 - 1\}$ represents the intensity level. We have concatenated the histogram values of the R, G, B channels as $[H(u, v)_L^R, H(u, v)_L^G, H(u, v)_L^B]$, and defined them as RGB-C where C stands for concatenation. After concatenation, there are 768 levels available (each channel has 256 levels). For the classification tools, 768 values of input features have been used.

Harlick Features

A GLCM is a two-dimensional matrix where every entry (i, j) represents the number of times a pixel of value i is adjacent to a pixel of value j , at a particular angle θ and distance d . Later this entire matrix is divided by the total number of comparisons.

$$G' = (a_{(i,j,\theta,d)}) \in R^{N_g, N_g}, \quad (9.12)$$

where $R \in 0, - - - - -, N_g - 1$. Let

$$G = \frac{G'}{\text{Total Number of comparisons}} \quad (9.13)$$

Table 9.1: Harlick Features mathematical representation

Feature Name	Feature Mathematical Representation	Feature Name	Feature Mathematical Representation
Angular Second Moment [2]	$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} [P_{(i,j)}]^2$	Contrast [2]	$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-j)^2 [P_{(i,j)}]$
Correlation [2]	$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{(i,j)[P_{(i,j)}] - \mu_i \mu_j}{\sigma_i \sigma_j}$	Variance [2]	$\frac{1}{N_g \times N_g} \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i - \mu_i)^2 p_{(i,j)} + (i - \mu_i)^2 p_{(i,j)}$
Inverse Difference Moment [2]	$\sum_i \sum_j \frac{1}{1+(i-j)^2} P_{(i,j)}$	Sum Average [2]	$\sum_{i=2}^{2N_g} i P_{X,Y}$
Sum Variance [2]	$\frac{1}{N_g \times N_g} \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} [ip_{i,j} + jp_{i,j}]$	Sum Entropy [2]	$-\sum_{i=2}^{2N_g} (i - f_s)^2 P_{X,Y}(i)$
Entropy [2]	$-\frac{1}{N_g \times N_g} \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p_{(i,j)} \log_2(p_{i,j})$	Difference Variance [2]	$\sum_{i=0}^{N_g-1} i^2 P_{X,Y}(i-1)$
Difference Entropy [2]	$-\sum_{i=2}^{2N_g} (i - f_s)^2 \log_2\{P_{X,Y}(i)\}$	Measure Of Correlation 1 [2]	$\frac{HXY - HXY1}{\max\{HX, HY\}}$
Maximum Correlation Coefficients [2]	$Q(i, j) = \sum_K \frac{P_{(i,k)} P_{(j,k)}}{P_X(i) P_Y(j)}$	Measure of Correlation 2 [2]	$((1 - \exp^{-2(HXY2 - HXY)})^{\frac{1}{2}})$

$$\mu_i = \sum_{j=1}^{N_g} i \sum_{j=1}^{N_g} p_{i,j}; \quad \mu_j = \sum_{i=1}^{N_g} j \sum_{i=1}^{N_g} p_{i,j}; \quad (9.14)$$

$$\sigma_i = \sum_{j=1}^{N_g} (i - \mu_i)^2 \sum_{j=1}^{N_g} p_{i,j}; \quad \sigma_j = \sum_{i=1}^{N_g} (j - \mu_j)^2 \sum_{i=1}^{N_g} p_{i,j} \quad (9.15)$$

The elements of G can be represented as $P_{i,j} = \frac{(a_{(i,j,\theta,d)})}{\text{Total Number of Comparisons}}$ where $0 \leq P_{i,j} \leq 1$. Based on the GLCM, statistical and structural features have been proposed by R. Harlick [8], who describes 14 different items of statistical information from the co-occurrence matrix, summarised in Table 9.1 [2].

Local Binary Pattern

The LBP was proposed by Ojala et al. [25], and is actually a rotation-invariant texture operator. Let $f_c(i_c, j_c)$ be a reference pixel position with strength measures as f_c . Each pixel within a radius r is assigned a value 0 or 1 depending on the value of the reference pixel $f_c(i_c, j_c)$. Suppose that there are w neighbouring pixels, then the LBP for the reference pixel is calculated based on the following formula:

$$f_c^{\text{LBP}} = \sum_{h=0}^{w-1} \mathcal{S}(f_h - f_c) 2^h \quad (9.16)$$

where

$$\mathcal{S}(f_h - f_c) = \begin{cases} 1, & \text{if } (f_h - f_c) \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

9.4 Comparison and Explanation of XGBoost

We have compared the XGBoost algorithm performance in respect to the Precision, Recall, F-measure and Sensitivity values with a few of the other available classifier models, in Tables 9.3, 9.4, 9.5 and 9.6 with Tamura, Histogram, LBP and Harlick features, respectively. XGBoost is a tree-based algorithm containing a few parameters. Along with

the other parameters we have fixed the parameters for the initial simulation as presented in Table 9.2.

Table 9.2: Initial Parameters for the XGBoost Algorithm

Parameter	Parameter Value
Learning Rate	0.1
Maximum depth of the tree	3
Number of Trees	100

Tables 9.3-9.7 summarise which classifier has the best performance in respect to the Precision, Recall, F-measure and Sensitivity values.

- The first column of Table 9.7 shows that the RIDOR algorithm provides the best Precision, and Random Forest provides the best F-measure, when we utilised Tamura features.
- The second column of Table 9.7 shows that Random Forest provides the best Precision and F-measure, whereas the IB1 and IB-K algorithms offer the best Specificity values, when we utilised Histogram information as a feature.
- The third column of Table 9.7 shows that the J-48 algorithm provides the best Precision, while the Random Forest algorithm provides the best F-measure, when we apply LBP features.
- The fourth column of Table 9.7 shows that the Random Forest algorithm has the best Precision, and both the IB1 and IB-K algorithms provide the best F-measure, when we employ Harlick features.

Table 9.3: Comparison of various classifiers with XGBoost algorithm using Tamura features

Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)
C4.5	92.90	90.80	91.80	84.90	AdaBoostM1	77.90	95.30	81.90	18.00
Decision Stamp	68.50	100.0	81.30	00.00	Simple Cart	92.60	97.00	91.60	84.00
ADTree	77.50	87.9	82.40	44.30	Bagging	94.40	94.30	94.30	87.70
Random Forest	94.80	97.30	96.00	88.30	Classifier Via Regression	94.70	91.40	93.00	89.00
XGBoost	92.60	95.40	94.00	83.40	Stacking C	68.50	100.0	81.30	49.90
Random Tree	88.50	88.40	88.50	75.30	END	92.90	90.80	91.80	84.90
Conjunctive Rule	68.50	100.0	81.30	00.00	Multi-Boost AB	68.50	100.0	81.30	00.00
JRip	87.70	88.70	88.20	72.90	Ordinal Class Classifier	92.90	90.80	91.80	84.90
Zero	68.50	100.0	81.30	72.90	RI Logit Boost	72.90	92.00	81.40	25.70
PART	93.20	87.60	90.30	86.10	Random Committee	92.70	96.60	94.6	83.50
ONE-R	72.50	83.40	77.60	30.90	Random Subspace	93.10	97.40	95.90	84.20
IB1	90.50	82.40	86.30	81.10	Non-linear SVM	68.50	100.0	81.30	00.00
IB-K	90.50	82.40	86.30	81.10	Hyper Pipes	68.70	99.90	81.40	80.00
BayesNet	77.80	67.30	72.20	58.30	RIDOR	96.60	87.70	91.90	93.20
Naive Bias Updateable	77.00	58.80	66.50	62.00	J-48 Graft	92.90	95.10	92.00	84.80
Naive Bias	77.00	58.50	66.50	62.00	Logit Boost	74.60	94.40	83.30	30.10
Dragging	68.90	99.20	81.30	02.60	Grading	68.50	100.0	81.30	00.00
Stacking	68.50	100.0	81.30	00.00	VFI	96.50	05.10	09.70	99.60

Table 9.4: Comparison of various classifiers with XGBoost algorithm using Histogram features

Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)
C4.5	92.10	91.80	92.00	83.00	AdaBoostM1	85.5	91.80	88.50	66.00
Decision Stamp	77.70	85.40	85.40	40.60	Simple Cart	91.10	93.80	92.40	80.10
ADTree	87.40	90.10	88.80	71.80	Bagging	92.80	95.50	94.10	83.90
Random Forest	94.20	97.02	95.70	86.90	Classifier Via Regression	91.90	94.10	93.00	82.00
XGBoost	92.10	95.01	93.60	82.30	Stacking C	68.50	100.0	81.30	00.00
Random Tree	91.40	91.80	91.60	81.20	END	92.10	91.80	92.00	83.00
Conjunctive Rule	80.90	92.20	86.20	52.60	Multi-Boost AB	77.70	95.20	85.60	40.06
JRip	92.00	93.50	92.70	82.20	Ordinal Class Classifier	92.10	91.80	92.00	83.00
Zero	68.50	100.0	81.30	00.00	RI Logit Boost	86.00	90.00	88.10	68.70
PART	91.50	92.80	92.10	81.10	Random Committee	93.20	97.40	98.00	84.50
ONE-R	76.90	89.60	82.80	41.50	Random Subspace	92.90	95.70	94.30	84.00
IB1	95.90	95.00	95.40	91.10	Non-linear SVM	93.20	97.40	95.20	84.50
IB-K	95.90	95.00	95.40	91.10	Hyper Pipes	70.50	99.10	82.40	09.80
BayesNet	88.00	83.60	85.70	75.20	RIDOR	88.30	95.40	91.70	72.50
Naive Bias Updateable	74.00	96.00	83.70	26.60	J-48 Graft	89.20	89.30	89.30	85.70
Naive Bias	74.00	96.20	37.00	26.60	Logit Boost	87.00	90.05	88.70	70.50
Dragging	86.70	94.50	90.40	68.50	Grading	68.50	100.0	81.30	00.0
Stacking	68.50	100.0	81.03	00.0	VFI	93.40	63.70	75.70	90.10

Table 9.5: Comparison of various classifiers with XGBoost algorithm using LBP features

Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)
C4.5	77.10	78.80	77.90	48.90	AdaBoostM1	74.60	85.60	79.70	36.40
Decision Stamp	74.70	82.90	78.60	39.00	Simple Cart	75.60	90.00	82.10	36.60
ADTree	74.30	93.10	82.70	29.40	Bagging	78.00	92.10	84.40	43.40
Random Forest	76.80	96.40	85.50	36.60	Classifier Via Regression	74.80	84.20	81.20	49.60
XGBoost	77.10	93.60	84.50	39.40	Stacking C	68.50	100.0	81.30	00.00
Random Tree	77.00	78.00	77.50	49.40	END	77.10	78.80	77.90	48.90
Conjunctive Rule	69.50	95.90	80.60	08.50	Multi-Boost AB	73.90	85.80	79.40	34.20
JRip	76.00	89.60	82.20	38.40	Ordinal Class Classifier	77.10	78.80	77.90	48.90
Zero	68.50	100.0	81.30	00.00	RI Logit Boost	73.40	90.60	81.50	27.50
PART	78.50	77.40	78.00	53.90	Random Committee	77.40	94.10	89.90	40.10
ONE-R	70.80	87.70	78.40	21.10	Random Subspace	76.00	94.80	84.40	35.00
IB1	83.40	82.30	82.80	64.20	Non-linear SVM	68.50	100.0	81.30	00.00
IB-K	83.40	82.30	82.80	64.20	Hyper Pipes	69.00	99.30	81.40	02.90
BayesNet	74.60	73.60	74.20	45.40	RIDOR	72.60	97.10	83.10	20.10
Naive Bias Updateable	73.20	80.90	76.80	35.30	J-48 Graft	89.20	89.30	89.30	85.70
Naive Bias	73.20	80.90	76.80	35.30	Logit Boost	74.60	91.60	82.30	32.02
Dragging	75.60	93.90	83.70	33.80	Grading	68.50	91.60	82.30	32.20
Stacking	68.50	100.0	81.30	00.00	VFI	81.70	09.10	16.40	95.60

Table 9.6: Comparison of various classifiers with XGBoost algorithm using Harlick features

Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Classifier	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)
C4.5	88.60	90.40	89.50	74.70	AdaBoostM1	80.80	92.60	86.30	52.00
Decision Stamp	78.60	75.90	77.20	54.40	Simple Cart	73.80	88.30	90.60	89.50
ADTree	83.70	93.10	87.60	57.70	Bagging	89.10	94.50	91.70	74.90
Random Forest	90.40	95.40	92.90	78.00	Classifier Via Regression	89.50	91.90	90.70	76.60
XGBoost	86.50	94.20	90.20	67.00	Stacking C	68.50	100.0	81.30	0.000
Random Tree	87.80	88.80	88.30	73.10	END	88.60	90.40	89.50	74.70
Conjunctive Rule	83.80	75.20	76.40	70.40	Multi Boost AB	77.20	92.40	84.10	40.70
JRip	87.10	92.20	89.60	83.60	Ordinal Class Classifier	88.60	90.40	89.50	74.70
Zero	68.50	100.0	81.30	00.00	RI Logit Boost	82.40	89.50	85.80	58.50
PART	85.00	91.30	88.00	65.00	Random Committee	89.5	95.20	92.20	75.70
ONE-R	74.50	84.40	79.10	60.70	Random Subspace	88.20	94.30	91.10	72.40
IB1	93.50	92.60	93.10	86.10	Non linear SVM	84.30	84.50	84.10	79.70
IB-K	93.50	92.60	93.10	86.10	Hyper Pipes	70.00	99.70	82.20	06.80
BayesNet	82.00	88.80	85.30	57.60	RIDOR	85.50	90.60	88.00	66.50
Naive Bias Updateable	72.10	90.50	80.30	23.90	J-48 Graft	88.80	91.30	90.00	74.80
Naive Bias	72.10	90.50	80.30	23.90	Logit Boost	81.90	92.20	86.70	55.50
Dragging	84.80	94.00	88.90	62.00	Grading	68.50	100.0	81.30	00.00
Stacking	68.50	100.0	81.30	00.00	VFI	70.00	97.00	82.30	07.00

Table 9.7: Overall best classifiers

	Precision	Recall	F-measure	Specificity
Tamura	(a) RIDOR	(a) Decision Stamp	(a) Random Forest	(a) VFI
		(b) Conjunctive Rule		
		(c) Zero (d) Stacking (e) Stacking C		
		(f) Multi-Boost AB (g) Non-linear SVM		
Histogram	(a) Random Forest	(h) Grading	(a) Random Forest	(a) IB1 (b) IB-K
		(a) Zero (b) Stacking (c) Stacking C		
		(d) Grading		
LBP	(a) J-48	(a) Zero (b) Stacking (c) Stacking C	(a) Random Forest	(a) VFI
		(c) Non-linear SVM		
Harlick	(a) Random Forest	(a) Zero (b) Stacking (c) Grading	(a) IB1 (b) IB-K	(a) Simple Cart

The above analysis shows that the XGBoost algorithm does not perform better than a few other existing classifiers. However, as a new classifier technique we have selected this classifier method for further analysis. Figures 9.5 a, b, c and d show the ROC curves for the LBP, Tamura, Histogram and Harlick features, respectively, when we utilised the XGBoost algorithm.

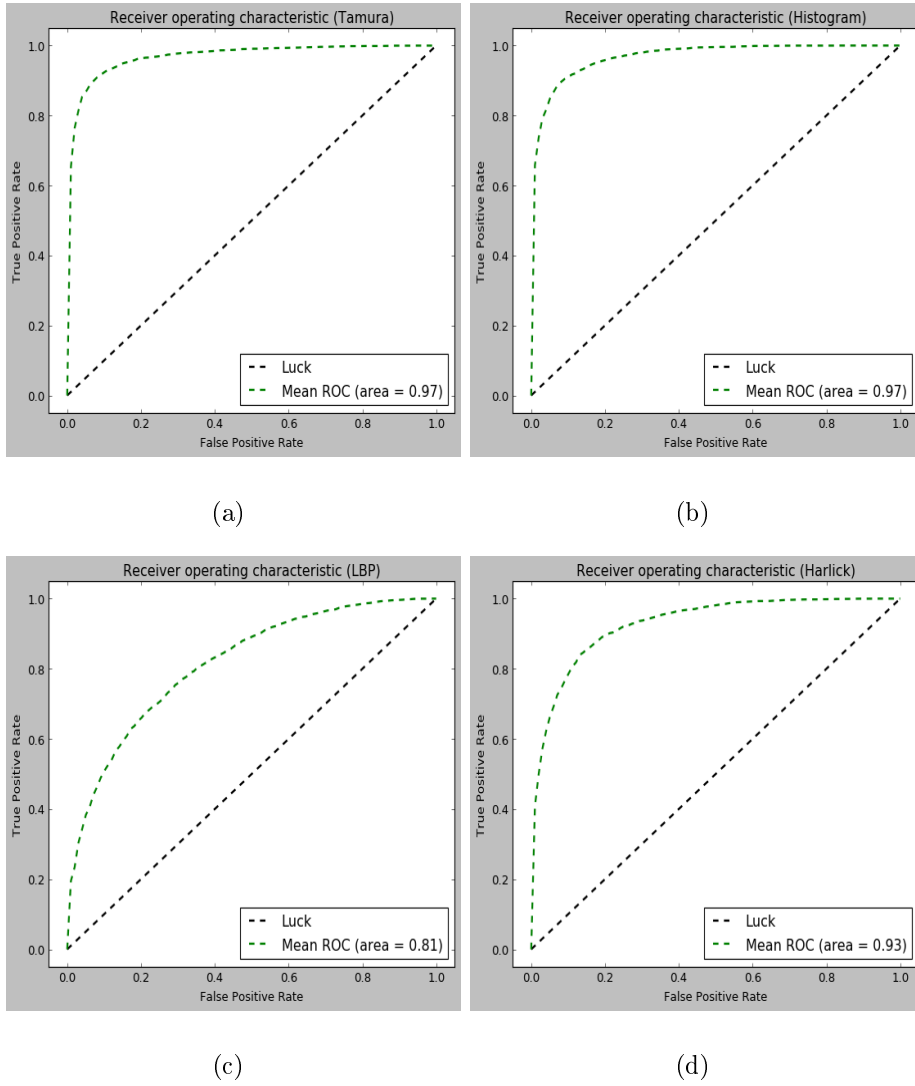


Figure 9.5: (a), (b), (c) and (d) show the ROC curves for the LBP, Tamura, Histogram and Harlick features, respectively, when we utilise the XGBoost algorithm.

Table 9.8 documents the accuracy performance when we utilise Tamura, Histogram,

LBP and Harlick features, respectively, along with the time required to construct the XGBoost model for the classification. LBP features give the worst Accuracy performance of 76.60% while Tamura and Histogram features give almost the same (91.67% and 91.14%) performance. However, in the case of the model construction time, when we utilise Tamura features it required 03.17s which was almost 20 times as small as the time required when we use Histogram features.

Table 9.8: Accuracy and Model Construction Time

Feature Name	Average Accuracy %	Elapsed Time (s)
LBP	76.60	48.16
Tamura	91.67	03.17
Histogram	91.14	61.39
Harlick	85.98	06.07

Tables 9.3, 9.4, 9.5, 9.6 and 9.8 and Figure 9.5 show that when we only consider the XGBoost algorithm, Tamura features function better than the LBP, Histogram and Harlick features for the Precision, Recall, F-measure and Sensitivity measuring criteria. As the Tamura features provide the best performance when we utilised the XGBoost algorithm, the following subsection tries to examine the XGBoost algorithm performance based on Tamura features while varying a few of the parameters of the XGBoost algorithm.

9.4.1 Performance based on XGBoost algorithm along with Tamura Features

Along with the other parameters, the Number of Trees, depth of trees and learning rate play vital roles in the XGBoost classifier performance. Figure 9.6a illustrates the per-

formance of the classifier (for Tamura) in respect to the Log Loss, Number of Trees and depth of tree. We have varied the Number of trees from 0 to 2000 keeping the depth of tree at 1, 3, 5, 7 and 9. The best performance in respect to the Log Loss is achieved when the Number of Trees is 1050 and the depth of trees is 5.

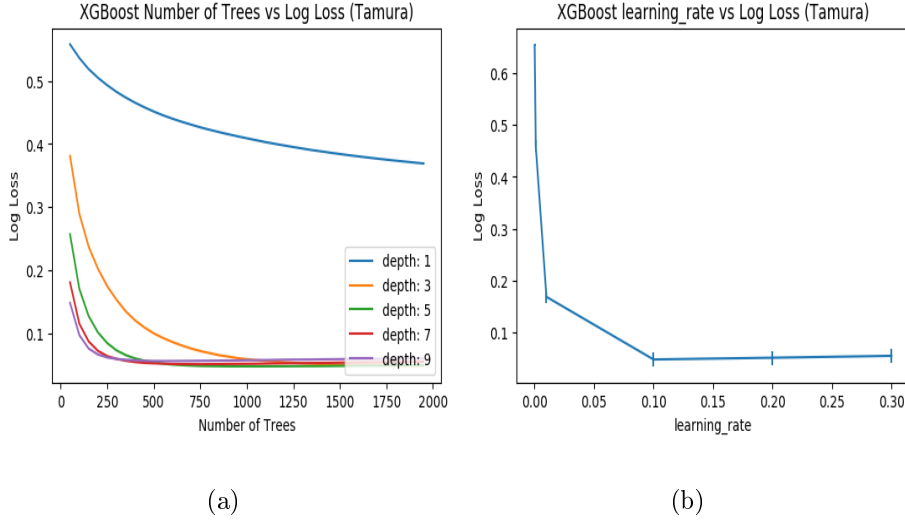


Figure 9.6: (a) shows the Log Loss performance against the Number of Trees along with the depth of tree. (b) shows the Log Loss value for different learning rates when the depth of tree is 5 and the Number of Trees is 1050.

Keeping the Number of Trees at 1050 and the depth of tree at 5, we changed the learning rate to find the best Log Loss value. Figure 9.6b shows that the Log Loss value is almost constant after the learning rate reaches 0.1. Based on Figure 9.6 a and b we have selected the values of the Number of Trees, depth of tree and the learning rate equal to 5, 1050 and 0.10, respectively, and recorded an Accuracy of 98.20% with a required time for the model construction of 44.07 s.

Figure 9.7a depicts the corresponding ROC curve. After selecting the values of the depth of tree, Number of Trees and learning rate, the model achieved 98.22% Accuracy, however, the computational time increased. Proper Feature-Selection can improve the computational time.

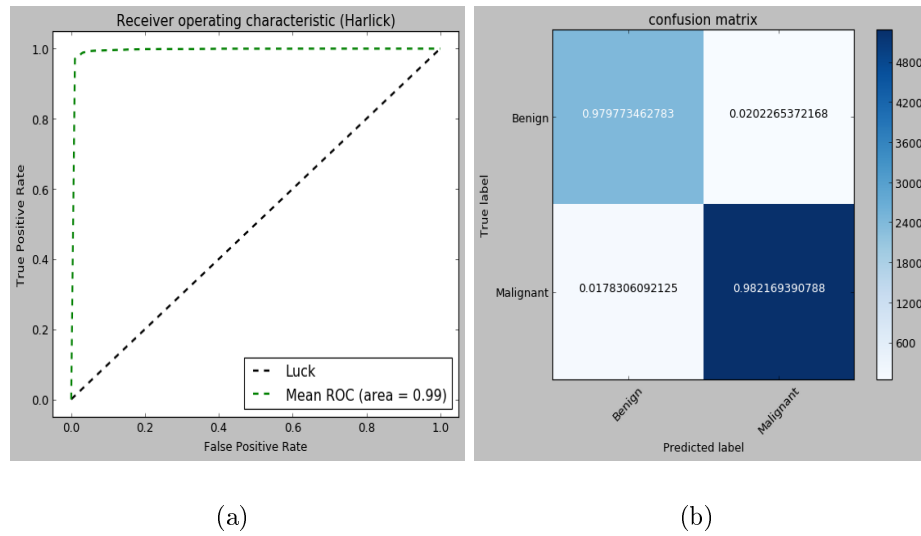


Figure 9.7: (a) shows the ROC curve and (b) represents the Confusion Matrix when the depth of tree is 5, Number of Trees is 1050, and learning rate is 0.10

Table 9.9: Average Recall/Accuracy/Precision/Specificity when depth of tree is 5 and Number of Trees is 1050

Recall%	Accuracy%	Precision%	F-measure%	Specificity%	Elapsed Time
97.90	98.22	98.14	98.06	98.64 s	44.07 s

9.5 Feature-Selection Methodology

Feature-Selection for the supervisor learning is always an important issue for a real-life classification problem. In a few cases the dataset may contain irrelevant information, and Feature-Selection can identify that irrelevant information. Overall a Feature-Selection method has the following important quantities to keep low

- dataset dimensionality
- computational complexity
- computational time.

Since all the features do not have the same importance for classification purposes we have utilised two Feature-Selection methods: a) Filter and b) Wrapper to find the best feature set for the classifier.

9.5.1 Filter

The filter method is applied before the classifier, and works as a pre-processing step to determine suitable features for the classifier. The filter method provides score values, and based on this score value a particular feature set is selected. A few Feature-Selection methods are available, however, we have utilised the following filter methods for our analysis:

- Chi-Square (CS) Method: The Chi-Square method establishes the likelihood of correlation using the frequency distribution

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (9.17)$$

where O_i represents the observed distribution and E_i represents the expected distribution. Features which produce greater χ^2 values have more importance for the data classification. Table 9.10 shows that $\mathcal{R}_{2\mathcal{R}}$ was the most important feature when we applied the CS method for Feature-Selection.

- Relief (RE) Method: Lira and Randell proposed the Relief algorithm, where each feature is weighted between the values of -1 and +1. A weight towards +1 is more significant for the classification.
- Fisher Score (FS) Method: Fisher's Discriminant Ratio for the binary-class classification can be written as

$$\mathcal{F} = \frac{(\mu_B - \mu_M)^2}{\sigma_B^2 - \sigma_M^2} \quad (9.18)$$

where μ_B , μ_M , σ_B and σ_M represent the average and the standard deviation of the Benign and Malignant classes for any feature.

- Mutual Information (MI): Mutual Information (MI) calculates the common infor-

mation between two random variables, and can be defined as

$$\begin{aligned} \mathcal{MJ}(\mathcal{A}, \mathcal{B}) &= \mathcal{H}(\mathcal{A}) + \mathcal{H}(\mathcal{B}) - \mathcal{H}(\mathcal{A}, \mathcal{B}) \\ &= - \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a, b) \times \log_2 \frac{p(a, b)}{p(a)p(b)} \end{aligned} \quad (9.19)$$

From the information-theoretical perspective, $\mathcal{H}(\mathcal{A})$ is the Entropy, defined as

$$\mathcal{H}(\mathcal{A}) = - \sum_{a \in \mathcal{A}} P(\mathcal{A} = a) \times \log_2 P(\mathcal{A} = a) \quad (9.20)$$

All the above-mentioned filter methods produce a score value for each of the feature vectors. Based on this score value each filter algorithm prioritises the feature vectors. Let \mathcal{C}_{1S} , \mathcal{C}_{2S} , \mathcal{D}_{1S} , \mathcal{L}_S , \mathcal{R}_{1S} , \mathcal{R}_{2S} represent the Coarseness, Contrast, Directionality, Likelihood, Regularity and Roughness of the respective channel's S where $S \in \{\text{R}, \text{G}, \text{B}\}$. As we have extracted features from all the available three channels, the total number of features will be eighteen. Table 9.10 arranges the feature vectors in a descending order based on the score value gained from different Feature-Selection algorithms. As an example, the 3rd row of Table 9.10 arranges the feature vectors in a descending order for the CS method, where the \mathcal{R}_{2R} feature vector contains the highest score, 14.50, and the \mathcal{R}_{1B} feature vector contains the lowest priority score, 0.001. Similarly, rows 5, 6 and 7 of Table 9.10 arrange the feature vectors in a descending order for the FS, RE and MI methods, respectively.

Table 9.10: Feature Priority table for various Feature-Selection Methods

Data arranged in the order of the score (decreasing)																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CS	\mathcal{R}_{2R}	\mathcal{C}_{1R}	\mathcal{R}_{2G}	\mathcal{C}_{1G}	\mathcal{R}_{2B}	\mathcal{C}_{1B}	\mathcal{D}_{1R}	\mathcal{D}_{1B}	\mathcal{L}_R	\mathcal{L}_G	\mathcal{D}_{1G}	\mathcal{L}_B	\mathcal{C}_{2R}	\mathcal{R}_{1G}	\mathcal{C}_{2G}	\mathcal{R}_{1R}	\mathcal{C}_{2B}	\mathcal{R}_{1B}
	14.50	14.49	02.64	02.64	02.43	02.43	01.08	00.23	00.16	00.04	00.03	00.03	00.12	0.004	0.003	0.001	0.001	0.001
FS	\mathcal{L}_R	\mathcal{C}_{2R}	\mathcal{R}_{2R}	\mathcal{C}_{1R}	\mathcal{L}_B	\mathcal{L}_G	\mathcal{D}_{1R}	\mathcal{C}_{2G}	\mathcal{R}_{2G}	\mathcal{C}_{1G}	\mathcal{R}_{1G}	\mathcal{R}_{2B}	\mathcal{C}_{1B}	\mathcal{C}_{2B}	\mathcal{R}_{1R}	\mathcal{D}_{1B}	\mathcal{R}_{1B}	\mathcal{D}_{1G}
	0.097	0.037	0.032	0.032	0.020	0.013	0.009	0.007	0.007	0.007	0.006	0.005	0.005	0.003	0.002	0.002	0.001	0.003
Re	\mathcal{R}_{2R}	\mathcal{C}_{1R}	\mathcal{C}_{2B}	\mathcal{R}_{2B}	\mathcal{C}_{1B}	\mathcal{L}_R	\mathcal{C}_{2R}	\mathcal{C}_{1G}	\mathcal{R}_{2G}	\mathcal{C}_{2G}	\mathcal{L}_G	\mathcal{L}_B	\mathcal{R}_{1G}	\mathcal{R}_{1R}	\mathcal{R}_{1B}	\mathcal{D}_{1R}	\mathcal{D}_{1G}	\mathcal{D}_{1B}
	0.006	0.006	0.003	0.002	0.002	0.002	0.002	-0.000	-0.000	-0.001	-0.001	-0.001	-0.001	-0.001	-0.002	-0.003	-0.003	-0.004
MI	\mathcal{L}_R	\mathcal{R}_{2R}	\mathcal{C}_{2R}	\mathcal{L}_G	\mathcal{C}_{1R}	\mathcal{C}_{1G}	\mathcal{R}_{2G}	\mathcal{L}_B	\mathcal{C}_{2B}	\mathcal{R}_{1g}	\mathcal{C}_{2G}	\mathcal{R}_{2B}	\mathcal{D}_{1R}	\mathcal{R}_{1B}	\mathcal{D}_{1G}	\mathcal{R}_{1R}	\mathcal{C}_{1B}	\mathcal{D}_{1B}
	0.057	0.027	0.023	0.023	0.021	0.018	0.018	0.016	0.012	0.010	0.009	0.009	0.005	0.005	0.005	0.002	0.002	0.000

We investigated how the priority vectors obtained from different Feature-Selection algorithms behave in our model. To do so, we produced eighteen feature sets for each filter method. For instance, concerning the CS method, we produced eighteen feature sets where the first feature set $S_1^{\text{CS}} = \{\mathcal{R}_{2\text{R}}\}$ contains only the first priority vectors obtained by the Chi-square method, $S_2^{\text{CS}} = \{\mathcal{R}_{2\text{R}}, \mathcal{C}_{1\text{R}}\}$ contains the first two priority vectors obtained by the CS method, and similarly S_{18}^{CS} contains all the feature vectors. Similarly, S_t^{FS} , S_t^{RE} and S_t^{MI} represent the feature set for the FS, RE and MI methods, respectively, where $t = \{1, 2, 3, \dots, 18\}$. When we utilise the CS method, $\{\mathcal{R}_{2\text{R}}\}$ stands out. That is, the feature set S_1^{CS} of Table 9.11 contains this feature only. The second most prominent feature with the CS method is $\mathcal{C}_{1\text{R}}$; this second feature set of Table 9.11 contains both the features $\mathcal{C}_{1\text{R}}$ and $\mathcal{R}_{2\text{R}}$. We have summarised all these vector sets for these four methods in Tables 9.11, 9.12, 9.13 and 9.14.

Table 9.11: All the feature sets based on the Chi-Square (CS) Feature-Selection method

Name	Feature Set
S_1^{CS}	$\{\mathcal{R}_{2R}\}$
S_2^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}\}$
S_3^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}\}$
S_4^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}\}$
S_5^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}\}$
S_6^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}\}$
S_7^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}\}$
S_8^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}\}$
S_9^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R\}$
S_{10}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G\}$
S_{11}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}\}$
S_{12}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B\}$
S_{13}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}\}$
S_{14}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}, \mathcal{R}_{1G}\}$
S_{15}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}, \mathcal{R}_{1G}, \mathcal{C}_{2G}\}$
S_{16}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{1R}\}$
S_{17}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{1R}, \mathcal{C}_{2B}\}$
S_{18}^{CS}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1B}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{D}_{1G}, \mathcal{L}_B, \mathcal{C}_{2R}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{1B}\}$

Table 9.12: All the feature sets based on the Fisher Score (FS) Feature-Selection method

Name	Feature Set
S_1^{FS}	$\{\mathcal{L}_R\}$
S_2^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}\}$
S_3^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}\}$
S_4^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}\}$
S_5^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B\}$
S_6^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G\}$
S_7^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}\}$
S_8^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}\}$
S_9^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}\}$
S_{10}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}\}$
S_{11}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}\}$
S_{12}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}\}$
S_{13}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}\}$
S_{14}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{C}_{2B}\}$
S_{15}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{R}_{1R}\}$
S_{16}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{R}_{1R}, \mathcal{D}_{1B}\}$
S_{17}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{R}_{1R}, \mathcal{D}_{1B}, \mathcal{R}_{1B}\}$
S_{18}^{FS}	$\{\mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{L}_B, \mathcal{L}_G, \mathcal{D}_{1R}, \mathcal{C}_{2G}, \mathcal{R}_{2G}, \mathcal{C}_{1G}, \mathcal{R}_{1G}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{R}_{1R}, \mathcal{D}_{1B}, \mathcal{R}_{1B}, \mathcal{D}_{1G}\}$

Table 9.13: All the feature sets based on the Relief (RE) Feature-Selection method

Name	Feature Set
S_1^{RE}	$\{\mathcal{R}_{2R}\}$
S_2^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}\}$
S_3^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}\}$
S_4^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}\}$
S_5^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}\}$
S_6^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R\}$
S_7^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}\}$
S_8^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}\}$
S_9^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}\}$
S_{10}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}\}$
S_{11}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G\}$
S_{12}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B\}$
S_{13}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}\}$
S_{14}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}, \mathcal{R}_{1R}\}$
S_{15}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}, \mathcal{R}_{1R}, \mathcal{R}_{1B}\}$
S_{16}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}, \mathcal{R}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1R}\}$
S_{17}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}, \mathcal{R}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1G}, \}$
S_{18}^{RE}	$\{\mathcal{R}_{2R}, \mathcal{C}_{1R}, \mathcal{C}_{2B}, \mathcal{R}_{2B}, \mathcal{C}_{1B}, \mathcal{L}_R, \mathcal{C}_{2R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{L}_B, \mathcal{R}_{1G}, \mathcal{R}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1R}, \mathcal{D}_{1G}, \mathcal{D}_{1B}\}$

Table 9.14: All the feature sets based on the Mutual Information (MI) Feature-Selection method

Name	Feature Set
S_1^{MI}	$\{\mathcal{L}_R\}$
S_2^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}\}$
S_3^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}\}$
S_4^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G\}$
S_5^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}\}$
S_6^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}\}$
S_7^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}\}$
S_8^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B\}$
S_9^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}\}$
S_{10}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}\}$
S_{11}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}\}$
S_{12}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}\}$
S_{13}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}\}$
S_{14}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}, \mathcal{R}_{1B}\}$
S_{15}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1G}\}$
S_{16}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1G}, \mathcal{R}_{1R}\}$
S_{17}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1G}, \mathcal{R}_{1R}, \mathcal{C}_{1B}\}$
S_{18}^{MI}	$\{\mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2R}, \mathcal{L}_G, \mathcal{C}_{1R}, \mathcal{C}_{1G}, \mathcal{R}_{2G}, \mathcal{L}_B, \mathcal{C}_{2B}, \mathcal{R}_{1G}, \mathcal{C}_{2G}, \mathcal{R}_{2B}, \mathcal{D}_{1R}, \mathcal{R}_{1B}, \mathcal{D}_{1G}, \mathcal{R}_{1R}, \mathcal{C}_{1B}, \mathcal{D}_{1B}\}$

9.5.2 Wrapper

The wrapper method depends on the classifier model. A general wrapper method can be described as follows:

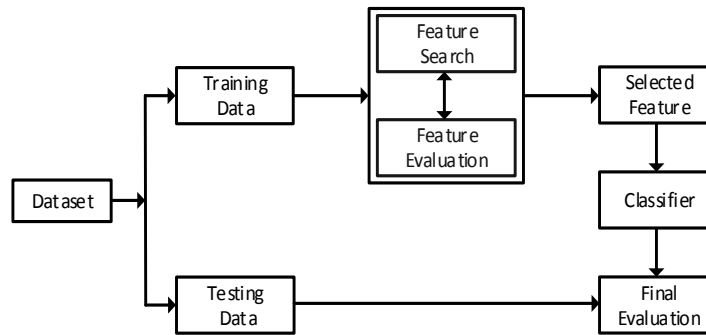


Figure 9.8: A wrapper method for Feature-Selection

- First Step: A set of features is selected from the training dataset.
- Second Step: Then the selected classifier performance is evaluated using that particular training feature set.
- The first and second steps continue until a desired performance criterion has been achieved.
- Those feature sets are selected which achieve the desired performance criterion.

Different wrapper Feature-Selection methods are available; from these we have executed the following two Feature-Selection operations:

Algorithm 4 Sequential Forward Selection Algorithm (SFS) [260]

-
- 1: Let the feature set be $\mathbf{x} = \{\mathbf{x}_j\}$ where $j = \{1, \dots, p\}$.
 - 2: Objective function $J_j(\cdot)$
 - 3: Let the empty feature set be $S_0 = \{\phi\}$
 - 4: **for** ($j=0$ to $p-1$) **do**
 - 5: Select the next-best feature as $\mathbf{x}_j^+ = \arg \max_{\mathbf{x} \notin S_j} J_j\{(S_j + \mathbf{x})\}$
 - 6: Update $S_{j+1} = S_j + \mathbf{x}_j^+$
 - 7: **end for**
-

Algorithm 5 Sequential Backward Selection Algorithm (SBS) [260]

-
- 1: Let the feature set be $\mathbf{x} = \{\mathbf{x}_i\}$ where $i = \{1, \dots, p\}$.
 - 2: Objective function $J_i(\cdot)$
 - 3: Let the full feature set be $S_p = \{\mathbf{X}\}$
 - 4: **for** ($i=p$ to 2) **do**
 - 5: Remove the next-worse feature as $\mathbf{x}_i^- = \arg \max_{\mathbf{x} \in S_i} J_i\{(S_i - \mathbf{x})\}$
 - 6: Update $S_{i-1} = S_i - \mathbf{x}_i^-$
 - 7: **end for**
-

We have determined the feature sets using the SFS and SBS algorithms and present those feature sets in Tables 9.15 and 9.16. Figure 9.9 (a) shows the Specificity performance for the different feature sets S_t^T , where $T = \{\text{CS, FS, RE, MI, SFS, SBS}\}$ and the value of $t = \{1, 2, 3, \dots, 18\}$. At $t = 1$, for all the feature set S_1^T the Specificity values are low, which means that for $S_1^{\text{CS}}, S_1^{\text{FS}}, S_1^{\text{RE}}, S_1^{\text{MI}}, S_1^{\text{SFS}}, S_1^{\text{SBS}}$ all the Specificity values are very low. When $t = 2$ or $t = 3$, the CS, FS, RE, MI algorithms provide less than 50 percent Specificity. However, both the feature sets $S_2^{\text{SFS}}, S_2^{\text{SBS}}$ provide very impressive Specificity values which touch around 92.00% and 93.00%, respectively. When $t \geq 2$, the values of the Specificity for the algorithms SBS and SFS never go below about 92.00%. At $t = 3, 4$, the Specificity values for the CFS algorithm demonstrate a slightly downward trend, however, after

Table 9.15: All the feature sets based on the Forward Feature-Selection method

Name	Feature Set
S_1^{SFS}	$\{\mathcal{L}_R\}$
S_2^{SFS}	$\{\mathcal{L}_R, \mathcal{L}_B\}$
S_3^{SFS}	$\{\mathcal{L}_R, \mathcal{L}_G, \mathcal{L}_B\}$
S_4^{SFS}	$\{\mathcal{L}_R, \mathcal{L}_G, \mathcal{C}_{2B}, \mathcal{L}_B\}$
S_5^{SFS}	$\{\mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{L}_G, \mathcal{C}_{2B}, \mathcal{L}_B\}$
S_6^{SFS}	$\{\mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{C}_{2B}, \mathcal{L}_B\}$
S_7^{SFS}	$\{\mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_8^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_9^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{2G}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_{10}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_{11}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_{12}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_{13}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{2B}\}$
S_{14}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{1B}, \mathcal{R}_{2B}\}$
S_{15}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{1R}, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{1B}, \mathcal{R}_{2B}\}$
S_{16}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{1R}, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{D}_{1G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{L}_B, \mathcal{R}_{1B}, \mathcal{R}_{2B}\}$
S_{17}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{R}_{1R}, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{D}_{1G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{D}_{1B}, \mathcal{L}_B, \mathcal{R}_{1B}, \mathcal{R}_{2B}\}$
S_{18}^{SFS}	$\{\mathcal{C}_{1R}, \mathcal{C}_{2R}, \mathcal{D}_{1R}, \mathcal{L}_R, \mathcal{R}_{1R}, \mathcal{R}_{2R}, \mathcal{C}_{1G}, \mathcal{C}_{2G}, \mathcal{D}_{1G}, \mathcal{L}_G, \mathcal{R}_{1G}, \mathcal{R}_{2G}, \mathcal{C}_{1B}, \mathcal{C}_{2B}, \mathcal{D}_{1B}, \mathcal{L}_B, \mathcal{R}_{1B}, \mathcal{R}_{2B}\}$

$t \geq 4$, the Specificity for the SFS algorithm never falls and always has an upward trend. Specifically, after $t \geq 8$ the Specificity values remain almost constant at around 88.50%. For the SBS algorithm, where $t = 3 - 5$ the Specificity algorithm shows a slightly up and down performance, however, in this period the Specificity never goes below around 91.00%. After $t \geq 8$, the Specificity value remains almost constant at 98.50%. For the cases CS, FS and MI, the Specificity values reveal a mediocre performance. Interestingly, when $t \geq 12$ the Specificity remains almost the same irrespective of the algorithm; this indicates that when $t \geq 12$, irrespective of the algorithm, almost 99.00% of Benign images are classified as Benign and only 1.00% have been misclassified as Malignant.

Figure 9.9(b) shows the Recall values for all the available feature sets for all the algorithms. For all the Feature-Selection sets and all the algorithms the Recall values never go below 81.00%. At $t = 1$, the Recall values for all the Feature-Selection algorithms are clustered around 88.00% to 90.00%. From $t = 2 - 6$, the CS, FS, and RE Feature-Selection algorithms show slightly lower Recall values than the initial ones. At $t = 2$, the RE, SFS, and SBS algorithms show a quite impressive Recall value of around 98.50%, which is maintained virtually throughout the whole period. The Recall data shows that never more than around 21% of the Malignant data have been misclassified as Benign for all the algorithms and all the feature sets. This figure also shows that, when $t \geq 12$, irrespective of the algorithm, around 98.00% of the Malignant images have been perfectly classified as Malignant images using any Feature-Selection algorithm.

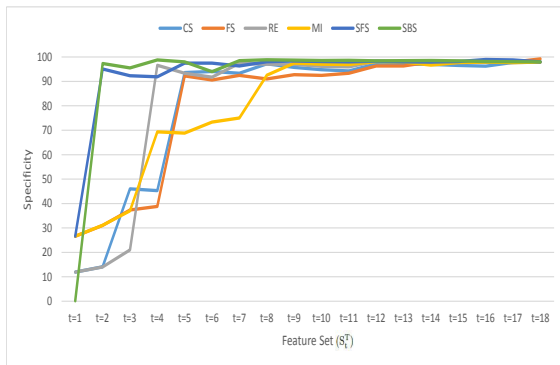
Figure 9.9 (c) depicts the Precision values for all the available feature sets for all the algorithms. For all the available Feature-Selection algorithms and their corresponding feature sets, the Precision values never go below around 67.50%. Initially, up to $t = 5$, the CS, FS and MI algorithms show quite poor Precision values. At $t = 5$, both the SFS and SBS algorithms provide very good Precision values which almost reach 98.75%. For $t = 2$ and 3, the Precision values for the SFS and SBS algorithms show a slightly

worse performance than for $t = 2$. For $t \geq 5$, the Precision values remain almost constant for both the SFS and SBS Feature-Selection algorithms, at around 99.55%. Initially, the Precision values from the RE Feature-Selection algorithm have poor performance, however, as the value of t rises, Precision values for the RE algorithm also perform in a similar way as for the SFS and SBS algorithms. For the Precision values, when $t \geq 12$ all the algorithms give almost the same performance, at around 99.50%.

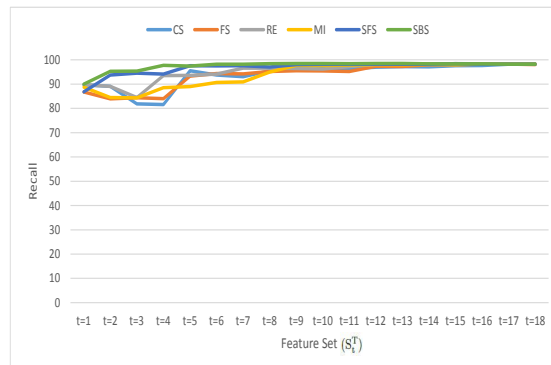
Figure 9.9 (d) depicts the F-measure values for all the feature sets comprising all the available Feature-Selection algorithms. For all the feature sets the SBS algorithm gives better F-measure values than the other algorithms. At $t = 2$, both the SFS and SBS algorithms offer very impressive F-measure values of around 98.00%. For $t = 1$ and $t = 3$ the CS, FS, RE and MI Feature-Selection algorithms show almost the same F-measure values of around 78.00%. When $t \geq 12$, the F-measure values remain constant for all the available sets and Feature-Selection algorithms.

Figure 9.9 (e) shows the Accuracy values for all the available feature sets for every available Feature-Selection algorithm. When $t = 1$ almost all the feature sets provide very poor Accuracy performance of around 66.00%. For $t = 2$, the SBS and SFS algorithms provide very good Accuracy performance, in fact the SBS algorithm shows slightly better performance than the SFS algorithm. For $t = 3 - 4$ the Accuracy for the SFS Feature-Selection algorithm remains the same as that for $t = 2$. For $t = 5 - 18$, both the SFS and SBS algorithms provide almost the same constant performance of almost 99.00%. For $t = 2 - 3$ the poor Accuracy value remains for the CS, FS, RE and MI algorithms. After $t = 2$, the accuracy values suddenly increase for the RE algorithm and the trend of increasing Accuracy continues.

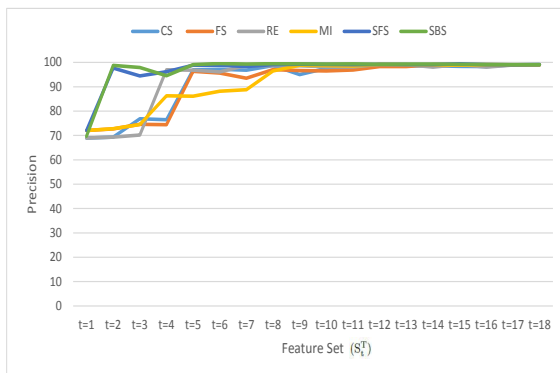
Figure 9.9 (e) illustrates the model construction time for the different feature sets. From $t = 1 - 4$, the model construction time is the same. As the value of t also represents the cardinality of the feature set, so when the cardinality of the feature set increases



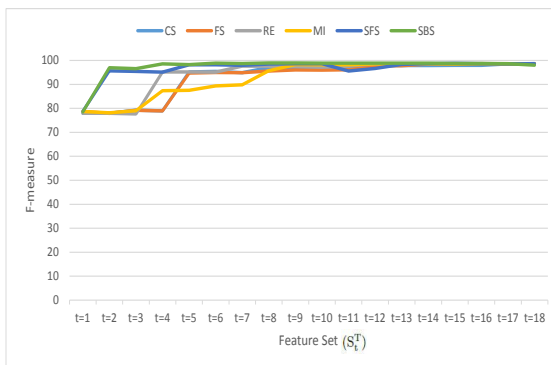
(a)



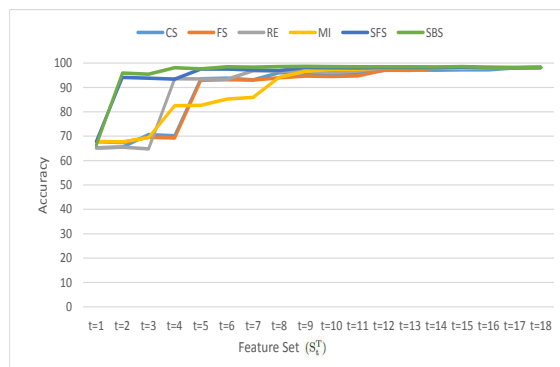
(b)



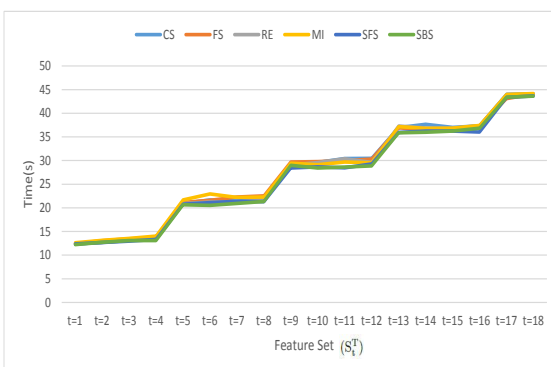
(c)



(d)



(e)



(f)

Figure 9.9: (a), (b), (c), (d), (e) and (f) show the Specificity, Recall, Precision, F-measure, Accuracy and model construction time for different feature sets based on different Feature-Selection algorithms.

from 1 to 4, the model takes almost the same time for construction, i.e. around 12 to 13 seconds. For $t = 5 - 8$, that is when the cardinality of the feature set varies from 5 to 8, the model construction time also remains constant, at around 20 seconds. For the feature-set cardinality 9 to 12 the model construction time remains almost 29 s. When the feature set contains 14, 15 and 16 features the model construction time is around 36.00 s. When the feature set contains 17 and 18 features the model construction time is approximately 42.00 s.

From the above discussion it is clear that if we take two features the best performance is given by the SBS algorithm and the feature set will be $S_2^{\text{SBS}} = \{\mathcal{C}_{2R}, \mathcal{C}_{2B}\}$. Here \mathcal{C}_{2R} and \mathcal{C}_{2B} represent the contrast of the red and blue channels, respectively. While $t = 2$, the SFS algorithm gives a slightly less accurate performance than the SBS algorithm, and the corresponding feature set is $S_2^{\text{SFS}} = \{\mathcal{L}_R, \mathcal{L}_B\}$. Here \mathcal{L}_R represents the Line-likeness for the red channel, and \mathcal{L}_B represents the Line-likeness for the blue channel. So feature sets S_2^{SBS} and S_2^{SFS} select different feature sets. For $t = 4$, the SBS algorithm gives almost 98.00% Accuracy and the elements of its feature set are $S_4^{\text{SBS}} = \{\mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2B}, \mathcal{L}_B\}$. Here $\{\mathcal{C}_{2R}, \mathcal{L}_R, \mathcal{C}_{2B}, \mathcal{L}_B\}$ represent the contrast of the RED channel, Line-likeness of the RED channel, contrast of the BLUE channel, Line-likeness of the BLUE Channel, respectively. For feature sets less than 12 ($t \leq 12$), in almost all situations the MI Feature-Selection algorithm performs worse than all the other available Feature-Selection algorithms.

9.6 Conclusion

In this chapter a set of Histopathological images has been classified into Benign and Malignant images utilising the state-of-the-art XG-Boost algorithm, along with Tamura, Harlick, LBP and Histogram features individually; of these the Tamura features provide the best performance. After fine-tuning the parameters such as the Number of Trees

(1050), depth of trees (5) and learning rate (0.10), we achieved 98.22% Accuracy with almost 98.00% Recall, Precision, F-measure and Specificity values when we use Tamura features, which requires around 44.07 s. It is found that proper selection of the feature set, such as the Contrast and Line-likeness features of the red channel and the contrast and Line-likeness of the blue channel provide the best Accuracy. This is almost 98.00%, when we utilised the Sequential Backward Feature-Selection algorithm, which requires almost 16 seconds.

Chapter 10

Conclusion and Future Work

10.1 Conclusion

Women are suffering from serious invasive Breast-Cancer (BC). In statistical terms BC is the second most common cause of death to women after lung cancer. Early identification of BC and proper diagnosis can increase a person's life expectancy. Doctors, physicians and radiologists investigate BC through physical examination of the patient, and also their decision about the diagnosis largely depends on investigation of biomedical images. Among the different biomedical images, histopathological images provide the most meaningful information on the disease. For this reason, doctors are heavily reliant on histopathological images to perfectly identify the current status of the disease. However, these images are complex in nature and their interpretation is very much subjective, requiring extensive expertise to identify malignancy. In particular cases, a Computer-Aided Diagnosis (CAD) system based on Machine-Learning (ML) provides a suitable solution to doctors. With the help of the CAD system, they can also compare their own decisions. As doctors are human and fallible, this provides more reliable decisions, and this conjugates two layers of decisions to provide extra satisfaction to patients.

A few ML algorithms are available based on different mathematical structures, a concept which has been utilised for BC image analysis, especially image classification. As BC is the cause of death of thousands of women, the involvement of ML in breast-image classifier design always has a huge importance. Thousands of new histopathological images are regularly produced which need to be diagnosed by utilising existing algorithms and new ones. As time goes on, new mathematical concepts and techniques are introduced which are adapted and implemented for BC image classification. With advances being made in computational architecture and applied mathematics, the state-of-the-art ML algorithms such as Deep Neural Network (DNN) and the Extreme Gradient Boosting Algorithm (XGBoost) algorithm have recently provided a significant improvement in data analysis. Using their advantages and some enhanced modifications, this thesis describes work which has classified a set of histopathological BC images into Benign and Malignant classes.

A DNN has the ability to extract global features as well as maintain hierarchical information. However, object-oriented local textural and statistical features provide a significant amount of extra information. Along with utilising global features, this dissertation investigates how local features combined with global features perform histopathological BC image classification.

- In Chapter 3, the statistical and structural information of each image has been clustered in an unsupervised way, and this clustered image is fed to the DNN model for the image classification. As a classifier model we have utilised: a) Convolutional Neural Network (CNN) model, b) Long Short Term Memory (LSTM) model, and c) CNN-LSTM model for the image classification. In the classifier layer both a Softmax layer and an Support Vector Machine (SVM) layer have been utilised individually, with their performances compared.
- In Chapter 4, as a pre-processing step, we have utilised a Retinex filter and applied

a CNN model for image classification. As the model we have utilised a) CNN model b) Resnet Model and c) Min-Max model for this purpose.

- Chapter 5 provides a novel architecture where the two local features:

- Histogram
- Local Binary Pattern (LBP)

have been extracted and then, along with the raw images, these features have been provided as input to the CNN model for classification. In the model, two parallel branches of the CNN model have been created, where in each branch the input data have been scanned through by a 5×5 random kernel to produce 16 feature maps. Then those two branches are concatenated together, until at the final decision layers a few similar subsequent branches have been created. This architecture shows that, as a local feature, Histogram information provides better performance than LBP features.

- As we know, frequency-domain information is also significant. Utilising frequency-domain information such as a) Discrete Fourier Transform (DFT) b) Discrete Cosine Transform (DCT) we have classified a set of histopathological BC images in Chapter 6. Learning from scratch is always better than learning from a reference point. A Recurrent Neural Network (RNN) utilises the mathematical concept where the model always learns from a reference point. The LSTM and GRU (advanced engineering of RNN) method has served in BC histopathological image classification.
- Local features such as Histogram LBP, frequency-domain features, DCT and DFT contain a significant amount of information. In Chapter 7 we have extracted the above-mentioned features along with the CT, and fed those features to the novel CNN model for the histopathological image classification. In the CNN model we have utilised the Resnet model.

- In Chapter 8 we have classified a set of histopathological BC images employing an unsupervised Deep Belief Network (DBN), which is created by stacking four layers of RBM one after another, and lastly it is guided by unsupervised backpropagation. For this we have utilised the Scale Conjugate Gradient method and as input we have utilised a set of Tamura features.
- In Chapter 9 this dissertation gives an extensive analysis of how the XGBoost algorithm performs for histopathological BC image classification where Histogram, LBP, Harlick and Tamura features have been utilised. The finding of the XGBoost algorithm has also been compared with those of a few other available existing classifier algorithms. Among these four features Tamura features provide the best performance with reference to Accuracy, Precision, Sensitivity, Receiver Operating Characteristic (ROC), and F-Measure values.

In this chapter, we have performed a few Feature-Selection algorithms, such as a) Filter and b)Wrapper, for finding the most prominent features. It is found that the concatenated feature vectors created by Contrast and Line-likeness of the Red and Blue channels provide 98.00% Accuracy. This feature-vector set has been found when we utilised the Sequential Backward Selection (SBS) algorithm.

This thesis found that integration of both global and local features may enhance the performance of the Deep Learning method for histopathological breast-cancer image classification (for the BreakHis dataset). However with proper preprocessing of the data with a fine tuning of the model, our methods can be utilised for the classification of other biomedical images such as X-ray images or other histopathological images.

10.2 Future Research Directions

In the future, research can be conducted on the following topics:

- Utilising DNN models for automatic biomedical image classification always suffers due to the non-availability of labelled training data, which can be solved by:

1. Data augmentation: In general and according to common knowledge, the more training data is fed to the DNN model the better the performance that can be achieved. New data can be synthetically generated, which is known as data-augmentation. A few of the data-augmentation methods such as Cropping, Rotating, Flipping, etc., can be performed to produce new synthetic data. However, care should be taken to provide protection against the over-fitting problem.
2. Transfer Learning: In a DNN transfer learning is the process where the knowledge of one network is transferred to perform another task, where normally the weights are shared. The models which are mostly utilised as inputs for a transfer learning model so far are:
 - a) VGG Model by Oxford, b) Inception Model by Google, and c) ResNet Model by Microsoft.

However, it is not wise to directly utilise the weight of the existing model for solving new problems. The accuracy performance might be degraded due to a negative weight transferring where the source knowledge is different for the target task. To avoid this and to obtain a reliable output the newly created network requires fine tuning for adjustment of the weight.

- Ensemble techniques: Ensemble is a well-known method where the classifier provides a decision based on comparing a few of the available classifiers' model outputs. Ensemble methods for the CNN model are not explored in detail. Instead of creating only one CNN network a set of CNN models can be created and their results combined in an ensemble to get the output. A CNN model can also be ensembled

with other classifier models such as RF to provide output decisions.

- **Commercialisation:** All the models based on CNN and LSTM in this thesis have been developed utilising the TensorFlow or Keras platform in a desktop computer. As both the CNN and LSTM methods are computationally expensive, we have utilised a Graphical Processing Unit (GPU) at the back end. These models can be accessible through a mobile or a light device, and can be used as a mobile diagnostic system if we can adapt our system for a low-computational device. For this there is scope to extend this work to make it applicable to light mobile devices by providing a light CNN and LSTM model. Instead of performing the computational operation in a central computational system, the computation can be done in a cloud-based system.
- **Extreme Learning (EL):** EL is another variety of the NN model which can be considered as shallow deep learning. In future, EL and the CNN model can be combined to create a light CNN model.
- **Data Reduction Techniques:** In this dissertation we have investigated Feature-Selection techniques to determine the important features. A few data-reduction techniques are available such as: a) Principal Component Analysis (PCA), b) Linear Discriminant Analysis (LDA), c) Auto-encoder. They can all be used for further data reduction.

Chapter 11

List of Abbreviations

AI	Artificial Intelligence
ALBP	Average Local Binary Pattern
ANOVA	Analysis of Variance
ASM	Angular Second Moment
BC	Breast-Cancer
BI-RADS	Breast Imaging Reporting and Data System
BRIEF	Binary Robust Independent Elementary Features
BBLBP	Block Based Local Binary Pattern
BW	Bandwidth
CAD	Computer-Aided Diagnosis
CAT	Computer-Aided Tomography
CLBP	Completed Modeling of Local Binary Pattern
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CS	Chi-Square
CT	Curvelet Transform

DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DDSM	Digital Database for Screening Mammography
DNN	Deep Neural Network
DFT	Discrete Fourier Transform
DoE	Difference of Entropy
DoG	Difference of Gaussian
DoV	Difference of Variance
FNR	False Negative Rate
FCM	Fuzzy C-Means Clustering
FPR	False Positive Rate
FAST	Feature From Accelerated Test
FS	Fisher Score
GAN	Generative Adversial Network
GLCM	Gray-Level Co-Occurrence Matrix
GLRM	Grey-Level Run-Length Matrix
GRU	Gated Recurrent Unit
HBD	Hessian Blob Detector
HD	Harris Detector
ID	Input Dimension
KM	K-Means
LBP	Local Binary Pattern
LSTM	Long Short Term Memory
LPQ	Local Plane Quantisation
LoG	Laplacian of Gaussian
M.C.C	Matthews Correlation Coefficient

MSE	Mean-Square Error
MI	Mutual Information
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MS	Mean Shift
NB	Naive Bayes
NN	Neural Network
ORB	Oriented FAST and rotated BRIEF
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SBS	Sequential Backward Selecion
SFS	Sequential Forward Selecion
SIFT	Scale Invariant Feature Transform
SOM	Self Organising Map
SONAR	Sound Navigation and Ranging
SURF	Speeded-Up Robust Features Descriptor
SoA	Sum of Averages
SoE	Sum of Entropy
SoV	Sum of Variance
SSoV	Sum of Squares of Variance

TS	Time Steps
TNR	True Negative Rate
TPR	True Positive Rate
SUSAN	Smallest Univalued Segment Assimilating Nucleus
SVM	Support Vector Machine
VLAD	Grassmannian Vector of Local Aggregated Descriptor
XGBoost	Extreme Gradient Boosting

Bibliography

- [1] “<http://www.aihw.gov.au/acim-books/>,”
- [2] <https://canceraustralia.gov.au/affected-cancer/what-cancer/cancer-australia-statistics>.
- [3] J. E. Skandalakis, *Embryology and Anatomy of the Breast*, pp. 3–24. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.
- [4] R. A. Shampo, Marc A. Kyle, “Karl Theodore Dussik-Pioneer in Ultrasound,” in *Mayo Clinic Proceedings*, p. 1136, 1995.
- [5] O. Karatas and E. Toy, “Three-dimensional imaging techniques: A literature review,” *European Journal of Dentistry*, vol. 8, no. 1, pp. 132–140, 2014.
- [6] M. Lakrimi, A. M. Thomas, G. Hutton, M. Kruip, R. Slade, P. Davis, A. J. Johnstone, M. J. Longfield, H. Blakes, S. Calvert, M. Smith, and C. A. Marshall, “The principles and evolution of magnetic resonance imaging,” *Journal of Physics: Conference Series*, vol. 286, no. 1, pp. 012–016, 2011.
- [7] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “Breast cancer histopathological image classification using convolutional neural networks,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2560–2567, July 2016.

- [8] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, pp. 786–804, May 1979.
- [9] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, pp. 460–473, June 1978.
- [10] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.
- [11] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, pp. 147–151, 1988.
- [12] S. M. Smith and J. M. Brady, "Susan—a new approach to low level image processing," *International Journal of Computer Vision*, vol. 23, no. 1, pp. 45–78, 1997.
- [13] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, pp. 1508–1515 vol. 2, Oct. 2005.
- [14] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the 9th European Conference on Computer Vision - Volume Part I, ECCV'06*, (Berlin, Heidelberg), pp. 430–443, Springer-Verlag, 2006.
- [15] R. Lenz, "Rotation-invariant operators and scale-space filtering," *Pattern Recognition Letters*, vol. 6, no. 3, pp. 151–154, 1987.
- [16] R. Lakemond, S. Sridharan, and C. Fookes, "Hessian-based affine adaptation of salient local image features," *Journal of Mathematical Imaging and Vision*, vol. 44, no. 2, pp. 150–167, 2012.

- [17] T. Lindeberg, "Scale selection properties of generalized scale-space interest point detectors," *Journal of Mathematical Imaging and Vision*, vol. 46, no. 2, pp. 177–210, 2013.
- [18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [19] W. Yussof and M. Hitam, "Invariant gabor-based interest points detector under geometric transformation," *Digital Signal Processing*, vol. 25, pp. 190–197, 2014.
- [20] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.
- [21] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 2, pp. II–257–II–263 vol. 2, June 2003.
- [22] B. Zhang, Y. Jiao, Z. Ma, Y. Li, and J. Zhu, "An efficient image matching method using speed up robust features," in *2014 IEEE International Conference on Mechatronics and Automation*, pp. 553–558, Aug. 2014.
- [23] B. Karasfi, T. S. Hong, A. Jalalian, and D. Nakhaeinia, "Speedup robust features based unsupervised place recognition for assistive mobile robot," in *2011 International Conference on Pattern Analysis and Intelligence Robotics*, vol. 1, pp. 97–102, June 2011.
- [24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, pp. 346–359, June 2008.

- [25] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, July 2002.
- [26] T. Ojala, M. Pietikäinen, and T. Mäenpää, “A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification,” in *Proceedings of the Second International Conference on Advances in Pattern Recognition*, ICAPR ’01, (London, UK, UK), pp. 397–406, Springer-Verlag, 2001.
- [27] T. Ahonen, J. Matas, C. He, and M. Pietikäinen, *Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features*, pp. 61–70. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.
- [28] G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915–928, June 2007.
- [29] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, “Brief: Computing a local binary descriptor very fast,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1281–1298, July 2012.
- [30] D. Gong, S. Li, and Y. Xiang, “Face recognition using the weber local descriptor,” in *The First Asian Conference on Pattern Recognition*, pp. 589–592, Nov. 2011.
- [31] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao, “WLD: A robust local image descriptor,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1705–1720, Sept. 2010.
- [32] S. H. Davarpanah, F. Khalid, L. Nurliyana Abdullah, and M. Golchin, “A texture descriptor: Background local binary pattern (bglbp),” *Multimedia Tools and Applications*, vol. 75, no. 11, pp. 6549–6568, 2016.

- [33] M. Heikkilä, M. Pietikäinen, and C. Schmid, “Description of interest regions with center-symmetric local binary patterns,” in *Computer Vision, Graphics and Image Processing*, (Berlin, Heidelberg), pp. 58–69, Springer Berlin Heidelberg, 2006.
- [34] G. Xue, L. Song, J. Sun, and M. Wu, “Hybrid center-symmetric local pattern for dynamic background subtraction,” in *2011 IEEE International Conference on Multimedia and Expo*, pp. 1–6, July 2011.
- [35] H. Wu, N. Liu, X. Luo, J. Su, and L. Chen, “Real-time background subtraction-based video surveillance of people by integrating local texture patterns,” *Signal, Image and Video Processing*, vol. 8, no. 4, pp. 665–676, 2014.
- [36] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, “Extended local binary patterns for face recognition,” *Information Sciences*, vol. 358–359, pp. 56–72, 2016.
- [37] T. Mäenpää and M. Pietikäinen, “Classification with color and texture: jointly or separately?,” *Pattern Recognition*, vol. 37, no. 8, pp. 1629–1640, 2004.
- [38] G. Xue, J. Sun, and L. Song, “Dynamic background subtraction based on spatial extended center-symmetric local binary pattern,” in *2010 IEEE International Conference on Multimedia and Expo*, pp. 1050–1054, July 2010.
- [39] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, “Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1301–1306, June 2010.
- [40] C. Silva, T. Bouwmans, and C. Frélicot, “An extended center-symmetric local binary pattern for background modeling and subtraction in videos,” in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP, (VISIGRAPP 2015)*, pp. 395–402, 2015.

-
- [41] Y. Chen, L. Ling, and Q. Huang, "Classification of breast tumors in ultrasound using biclustering mining and neural network," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1787–1791, Oct. 2016.
 - [42] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.
 - [43] N. D. Marom, L. Rokach, and A. Shmilovici, "Using the confusion matrix for improving ensemble classifiers," in *Electrical and Electronics Engineers in Israel (IEEEI), 2010 IEEE 26th Convention of*, pp. 555–559, Nov. 2010.
 - [44] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, (Amsterdam, The Netherlands, The Netherlands), pp. 3–24, IOS Press, 2007.
 - [45] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, pp. 65–86, 1958.
 - [46] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359–366, July 1989.
 - [47] R. Hecht-Nielsen, "Neural networks for perception (vol. 2)," ch. Theory of the Back-propagation Neural Network, pp. 65–93, Orlando, FL, USA: Harcourt Brace & Co., 1992.

- [48] J. Li, J.-H. Cheng, J.-Y. Shi, and F. Huang, *Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement*, pp. 553–558. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
- [49] A. Dawson, R. Austin Jr., and D. Weinberg, “Nuclear grading of breast carcinoma by image analysis: Classification by multivariate and neural network analysis,” *American Journal of Clinical Pathology*, vol. 95, no. 4 SUPPL. 1, pp. S29–S37, 1991.
- [50] K. T. Rajakeerthana, C. Velayutham, and K. Thangavel, *Mammogram Image Classification Using Rough Neural Network*, pp. 133–138. New Delhi: Springer India, 2014.
- [51] V. Lessa and M. Marengoni, *Applying Artificial Neural Network for the Classification of Breast Cancer Using Infrared Thermographic Images*, pp. 429–438. Cham: Springer International Publishing, 2016.
- [52] S. Wan, H.-C. Lee, X. Huang, T. Xu, X. Zeng, Z. Zhang, Y. Sheikine, J. L. Connolly, J. G. Fujimoto, and C. Zhou, “Integrated local binary pattern texture features for classification of breast tissue imaged by optical coherence microscopy,” *Medical Image Analysis*, vol. 38, pp. 104–116, 2017.
- [53] S. M. Lima, A. G. Silva-Filho, and W. P. dos Santos, “Detection and classification of masses in mammographic images in a multi-kernel approach,” *Computer Methods and Programs in Biomedicine*, vol. 134, pp. 11–29, 2016.
- [54] C. Abirami, R. Harikumar, and S. R. S. Chakravarthy, “Performance analysis and detection of micro calcification in digital mammograms using wavelet features,” in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2327–2331, March 2016.

- [55] N. E. Atlas, A. Bybi, and H. Drissi, "Features fusion for characterizing inbreast-database masses," in *2016 International Conference on Electrical and Information Technologies (ICEIT)*, pp. 374–379, May 2016.
- [56] H. Alharbi, G. Falzon, and P. Kwan, "A novel feature reduction framework for digital mammogram image classification," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 221–225, Nov. 2015.
- [57] W. Peng, R. Mayorga, and E. Hussein, "An automated confirmatory system for analysis of mammograms," *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 134–144, 2016.
- [58] A. Jalalian, S. Mashohor, R. Mahmud, B. Karasfi, M. Iqbal Saripan, and A. R. Ramli, "Computer-assisted diagnosis system for breast cancer in computed tomography laser mammography (ctlm)," *Journal of Digital Imaging*, pp. 1–16, 2017.
- [59] H. Li, X. Meng, T. Wang, Y. Tang, and Y. Yin, "Breast masses in mammography classification with local contour features," *BioMedical Engineering OnLine*, vol. 16, no. 1, pp. 1–12, 2017.
- [60] D.-R. Chen, R.-F. Chang, and Y.-L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," *Radiology*, vol. 213, no. 2, pp. 407–412, 1999.
- [61] D.-R. Chen, R.-F. Chang, Y.-L. Huang, Y.-H. Chou, C.-M. Tiu, and P.-P. Tsai, "Texture analysis of breast tumors on sonograms," *Seminars in Ultrasound, CT and MRI*, vol. 21, no. 4, pp. 308–316, 2000.
- [62] D.-R. Chen, R.-F. Chang, W.-J. Kuo, M.-C. Chen, and Y.-L. Huang, "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks," *Ultrasound in Medicine and Biology*, vol. 28, no. 10, pp. 1301–1310, 2002.

- [63] S. Silva, M. Costa, W. Pereira, and C. Filho, “Breast tumor classification in ultrasound images using neural networks with improved generalization methods,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6321–6325, Aug. 2015.
- [64] I. Saritas, “Prediction of breast cancer using artificial neural networks,” *Journal of Medical Systems*, vol. 36, no. 5, pp. 2901–2907, 2012.
- [65] E. López-Meléndez, L. D. Lara-Rodríguez, E. López-Olazagastí, B. Sánchez-Rinza, and E. Tepichin-Rodríguez, “Bicad: Breast image computer aided diagnosis for standard birads 1 and 2 in calcifications,” in *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, pp. 190–195, Feb. 2012.
- [66] <https://github.com/BVLC/caffe>.
- [67] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *CoRR*, vol. abs/1408.5093, 2014.
- [68] <http://torch.ch/>.
- [69] <http://www.vlfeat.org/matconvnet/>.
- [70] A. Vedaldi and K. Lenc, “Matconvnet - convolutional neural networks for MATLAB,” *CoRR*, vol. abs/1412.4564, 2014.
- [71] <http://deeplearning.net/software/theano/>.
- [72] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, “Theano: A cpu and gpu math compiler in python,” in *Proceedings of the 9th Python in Science Conference*, pp. 3–10, 2010.

- [73] <https://www.tensorflow.org/>.
- [74] <https://github.com/Microsoft/CNTK>.
- [75] <https://keras.io>.
- [76] <https://github.com/ml4j>.
- [77] <http://ceit.aut.ac.ir/~keyvanrad/DeeBNet>.
- [78] M. A. Keyvanrad and M. M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented MATLAB toolbox (deebnet)," *CoRR*, vol. abs/1408.3264, 2014.
- [79] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.
- [80] C. Y. Wu, S.-C. B. Lo, M. T. Freedman, A. Hasegawa, R. A. Zuurbier, and S. K. Mun, "Classification of microcalcifications in radiographs of pathological specimen for the diagnosis of breast cancer," vol. 2167, pp. 630–641, 1994.
- [81] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 598–610, Oct. 1996.
- [82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

- [83] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.
- [84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, 2015.
- [85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, 2015.
- [86] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.
- [87] P. Fonseca, J. Mendoza, J. Wainer, J. Ferrer, J. Pinto, J. Guerrero, and B. Castaneda, "Automatic breast density classification using a convolutional neural network architecture search procedure," *Proc. SPIE*, vol. 9414, 2015.
- [88] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1299–1312, May 2016.
- [89] Y. Liu, S. Zhou, and Q. Chen, "Discriminative deep belief networks for visual data classification," *Pattern Recognition*, vol. 44, no. 10–11, pp. 2287–2296, 2011.
- [90] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Systems with Applications*, vol. 46, pp. 139–144, 2016.
- [91] J. Zhang, J. I. Silber, and M. A. Mazurowski, "Modeling false positive error making patterns in radiology trainees for improved mammography education," *Journal of Biomedical Informatics*, vol. 54, pp. 50–57, 2015.

- [92] S. Lo, H. Li, Y. Wang, L. Kinnard, and M. T. Freedman, "A multiple circular path convolution neural network system for detection of mammographic masses," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 150–158, Feb. 2002.
- [93] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016.
- [94] H. Su, F. Liu, Y. Xie, F. Xing, S. Meyyappan, and L. Yang, "Region segmentation in histopathological breast cancer images using deep convolutional neural network," vol. 2015-July, pp. 55–58, 2015.
- [95] K. Sharma and B. Preet, "Classification of mammogram images by using cnn classifier," *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, pp. 2743–2749, 2016.
- [96] A. Albayrak and G. Bilgin, "Mitosis detection using convolutional neural network based features," *CINTI 2016 - 17th IEEE International Symposium on Computational Intelligence and Informatics: Proceedings*, pp. 335–340, 2017.
- [97] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, 2016.
- [98] M. Zejmo, M. Kowal, J. Korbicz, and R. Monczak, "Classification of breast cancer cytological specimen using convolutional neural network," *Journal of Physics: Conference Series*, vol. 783, no. 1, 2017.
- [99] F. Jiang, H. Liu, S. Yu, and Y. Xie, "Breast mass lesion classification in mammograms by transfer learning," in *Proceedings of the 5th International Conference*

- on Bioinformatics and Computational Biology*, ICBCB '17, (New York, NY, USA), pp. 59–62, ACM, 2017.
- [100] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, N. Sugita, Y. Kawasumi, T. Ishibashi, and M. Yoshizawa, “Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis,” in *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 1382–1386, Sept. 2016.
- [101] Y. Qiu, Y. Wang, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng, “An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology,” *Proc. SPIE*, vol. 9785, pp. 97850X–97850X–6, 2016.
- [102] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, K. Cha, and M. A. Helvie, “Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis,” *Proc. SPIE*, vol. 9785, 2016.
- [103] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, “Large scale deep learning for computer aided detection of mammographic lesions,” *Medical Image Analysis*, vol. 35, pp. 303–312, 2017.
- [104] K. J. Geras, S. Wolfson, S. G. Kim, L. Moy, and K. Cho, “High-resolution breast cancer screening with multi-view deep convolutional neural networks,” *CoRR*, vol. abs/1703.07047, 2017.
- [105] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, pp. 81–106, Mar. 1986.

-
- [106] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
 - [107] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," pp. 310–314, Dec. 2016.
 - [108] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.
 - [109] L. Breiman, "Arcing classifiers," *The Annals of Statistics*, vol. 26, no. 3, pp. 801–824, 1998.
 - [110] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
 - [111] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.
 - [112] S. Beura, "Classification of mammogram using two-dimensional discrete orthonormal s-transform for breast cancer detection," *Healthcare Technology Letters*, vol. 2, pp. 46–51(5), April 2015.
 - [113] J. Diz, G. Marreiros, and A. Freitas, *Using Data Mining Techniques to Support Breast Cancer Diagnosis*, pp. 689–700. Cham: Springer International Publishing, 2015.
 - [114] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," *13th International Conference on Intelligent Systems Design and Applications*, pp. 121–125, Dec. 2013.

- [115] A. Paul, A. Dey, D. P. Mukherjee, J. Sivaswamy, and V. Tourani, *Regenerative Random Forest with Automatic Feature Selection to Detect Mitosis in Histopathological Breast Cancer Images*, pp. 94–102. Cham: Springer International Publishing, 2015.
- [116] Z. Chen, M. Berks, S. Astley, and C. Taylor, *Classification of Linear Structures in Mammograms Using Random Forests*, pp. 153–160. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
- [117] Y. Zhang, B. Zhang, and W. Lu, “Breast cancer classification from histological images with multiple features and random subspace classifier ensemble,” *AIP Conference Proceedings*, vol. 1371, no. 1, pp. 19–28, 2011.
- [118] S. P. Angayarkanni and N. B. Kamal, “MRI mammogram image classification using id3 algorithm,” in *IET Conference on Image Processing (IPR 2012)*, pp. 1–5, July 2012.
- [119] K. Wang, M. Dong, Z. Yang, Y. Guo, and Y. Ma, “Regions of micro-calcifications clusters detection based on new features from imbalance data in mammograms,” *Proc. SPIE*, vol. 10225, pp. 102252C–102252C–6, 2017.
- [120] D. Bruno, M. Nascimento, R. Ramos, V. Batista, L. Neves, and A. Martins, “LBP operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues,” *Expert Systems with Applications*, vol. 55, pp. 329–340, 2016.
- [121] C. Muramatsu, T. Hara, T. Endo, and H. Fujita, “Breast mass classification on mammograms using radial local ternary patterns,” *Computers in Biology and Medicine*, vol. 72, pp. 43–53, 2016.
- [122] M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, and K. Wang, “An efficient approach for automated mass segmentation and classification in mammograms,” *Journal of Digital Imaging*, vol. 28, no. 5, pp. 613–625, 2015.

- [123] G. Piantadosi, R. Fusco, A. Petrillo, M. Sansone, and C. Sansone, *LBP-TOP for Volume Lesion Classification in Breast DCE-MRI*, pp. 647–657. Cham: Springer International Publishing, 2015.
- [124] I. El-Naqa, Y. Yang, M. Wernick, N. Galatsanos, and R. Nishikawa, “A support vector machine approach for detection of microcalcifications,” *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.
- [125] R. Chang, W. Wu, W. Moon, and D. Chen, “Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis,” *Ultrasound in Medicine and Biology*, vol. 29, no. 5, pp. 679–686, 2003.
- [126] Y. Chu, L. Li, D. B. Goldgof, Y. Qui, and R. A. Clark, “Classification of masses on mammograms using support vector machine,” *Proc. SPIE*, vol. 5032, pp. 940–948, 2003.
- [127] B. K. Singh, K. Verma, A. Thoke, and J. S. Suri, “Risk stratification of 2d ultrasound-based breast lesions using hybrid feature selection in machine learning paradigm,” *Measurement*, vol. 105, pp. 146–157, 2017.
- [128] J. Ding, H. D. Cheng, J. Huang, J. Liu, and Y. Zhang, “Breast ultrasound image classification based on multiple-instance learning,” *Journal of Digital Imaging*, vol. 25, no. 5, pp. 620–627, 2012.
- [129] M. Pang, Y. Wang, and J. Li, “Dirichlet-based concentric circle feature transform for breast mass classification,” *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2016-January, pp. 272–277, 2016.
- [130] N. Mhala and S. Bhandari, “Improved approach towards classification of histopathology images using bag-of-features,” *International Conference on Signal and Information Processing, IConSIP*, 2017.

- [131] C. Hiba, Z. Hamid, and A. Omar, "An improved breast tissue density classification framework using bag of features model," *4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 405–409, Oct. 2016.
- [132] B. Malik, J. Klock, J. Wiskin, and M. Lenox, "Objective breast tissue image classification using quantitative transmission ultrasound tomography," *Scientific Reports*, vol. 6, no. 3, 2016. Art. ID 38857.
- [133] R. Chang, W. Wu, W. Moon, Y. Chou, and D. Chen, "Support vector machines for diagnosis of breast tumors on US images," *Academic Radiology*, vol. 10, no. 2, pp. 189–197, 2003.
- [134] C. Akbay, N. G. Gençer, and G. Gençer, "CAD for detection of microcalcification and classification in mammograms," in *2014 18th National Biomedical Engineering Meeting*, pp. 1–4, Oct. 2014.
- [135] J. Levman, T. Leung, P. Causer, D. Plewes, and A. L. Martel, "Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines," *IEEE Transactions on Medical Imaging*, vol. 27, pp. 688–696, May 2008.
- [136] L. Martins, E. Silva, A. Silva, A. Paiva, and M. Gattass, "Classification of breast masses in mammogram images using ripley's k function and support vector machine," *Machine Learning and Data Mining in Pattern Recognition: 5th International Conference, Leipzig, Germany*, pp. 784–794, July 2007.
- [137] Y. Zhang, S. Wang, G. Liu, and J. Yang, "Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional fourier transform," *Advances in Mechanical Engineering*, vol. 8, no. 2, pp. 1–11, 2016.
- [138] F. Shirazi and E. Rashedi, "Detection of cancer tumors in mammography images using support vector machine and mixed gravitational search algorithm," in *2016 1st*

- Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, pp. 98–101, March 2016.
- [139] M. Sewak, P. Vaidya, C. C. Chan, and Z.-H. Duan, “Svm approach to breast cancer classification,” in *Second International Multi-Symposium on Computer and Computational Sciences (IMSCCS 2007)*, pp. 32–37, Aug. 2007.
- [140] J. Dheeba and S. T. Selvi, “Classification of malignant and benign microcalcification using svm classifier,” in *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, pp. 686–690, March 2011.
- [141] M. Taheri, G. Hamer, S. H. Son, and S. Y. Shin, “Enhanced breast cancer classification with automatic thresholding using svm and harris corner detection,” in *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, RACS '16, (New York, NY, USA), pp. 56–60, ACM, 2016.
- [142] M. Tan, J. Pu, and B. Zheng, “Optimization of breast mass classification using sequential forward floating selection (sffs) and a support vector machine (svm) model,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 6, pp. 1005–1020, 2014.
- [143] S. Kavitha and K. Thyagarajan, *Features Based Mammogram Image Classification Using Weighted Feature Support Vector Machine*, pp. 320–329. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
- [144] L. Pat, W. Iba, and T. Kevin, “An analysis of bayesian classifiers,” in *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, pp. 223–228, AAAI Press, 1992.

- [145] A. Tosun, A. B. Bener, and S. Akbarinasaji, “A systematic literature review on the applications of bayesian networks to predict software quality,” *Software Quality Journal*, vol. 25, no. 1, pp. 273–305, 2017.
- [146] J. Grover, *A Literature Review of Bayes’ Theorem and Bayesian Belief Networks (BBN)*, pp. 11–27. New York, NY: Springer New York, 2013.
- [147] S. Butler, G. Webb, and R. Lewis, “A case study in feature invention for breast cancer diagnosis using x-ray scatter images,” *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 2903, pp. 677–685, 2003.
- [148] E. Fischer, J. Lo, and M. Markey, “Bayesian networks of bi-radsTM descriptors for breast lesion classification,” *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 26 IV, pp. 3031–3034, 2004.
- [149] D. Soria, J. Garibaldi, E. Biganzoli, and I. Ellis, “A comparison of three different methods for classification of breast cancer data,” *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, pp. 619–624, 2008.
- [150] <http://www.cs.waikato.ac.nz/ml/weka/>.
- [151] E. J. Kendall and M. T. Flynn, “Automated breast image classification using features from its discrete cosine transform,” *PLOS ONE*, vol. 9, pp. 1–8, March 2014.
- [152] V. Oleksyuk, F. Saleheen, D. F. Caroline, S. A. Pascarella, and C. H. Won, “Classification of breast masses using tactile imaging system and machine learning algorithms,” *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4, Dec. 2016.

- [153] F. Burling-Claridge, M. Iqbal, and M. Zhang, “Evolutionary algorithms for classification of mammographic densities using local binary patterns and statistical features,” in *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 3847–3854, July 2016.
- [154] U. Raghavendra, U. R. Acharya, H. Fujita, A. Gudigar, J. H. Tan, and S. Chokkadi, “Application of gabor wavelet and locality sensitive discriminant analysis for automated identification of breast cancer using digitized mammogram images,” *Applied Soft Computing*, vol. 46, pp. 151–161, 2016.
- [155] N. P. Pérez, M. A. G. López, A. Silva, and I. Ramos, “Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography,” *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 19–31, 2015.
- [156] G. Rashmi, A. Lekha, and N. Bawane, “Analysis of efficiency of classification and prediction algorithms (naïve bayes) for breast cancer dataset,” *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology, ICERECT 2015*, pp. 108–113, 2016.
- [157] G. Gatuha and T. Jiang, “Android based naive bayes probabilistic detection model for breast cancer and mobile cloud computing: Design and implementation,” *International Journal of Engineering Research in Africa*, vol. 21, pp. 197–208, 2016.
- [158] M. Benndorf, E. Kotter, M. Langer, C. Herda, Y. Wu, and E. Burnside, “Development of an online, publicly accessible naive bayesian decision support tool for mammographic mass lesions based on the american college of radiology ACR bi-rads lexicon,” *European Radiology*, vol. 25, no. 6, pp. 1768–1775, 2015.
- [159] V. Rodríguez-López and R. Cruz-Barbosa, “Improving bayesian networks breast mass diagnosis by using clinical data,” *Lecture Notes in Computer Science (includ-*

- ing subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 9116, pp. 292–301, 2015.
- [160] K. Nugroho, N. Setiawan, and T. Adji, “Cascade generalization for breast cancer detection,” *Proceedings - 2013 International Conference on Information Technology and Electrical Engineering: "Intelligent and Green Technologies for Sustainable Development"*, ICITEE 2013, pp. 57–61, 2013.
- [161] V. Rodríguez-López and R. Cruz-Barbosa, “On the breast mass diagnosis using bayesian networks,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8857, pp. 474–485, 2014.
- [162] S. Sivakumari, R. Praveena Priyadarsini, and P. Amudha, “Accuracy evaluation of c4.5 and naïve bayes classifiers using attribute ranking method,” *International Journal of Computational Intelligence Systems*, vol. 2, no. 1, pp. 60–68, 2009.
- [163] V. Rodríguez-López and R. Cruz-Barbosa, “Improving bayesian networks breast mass diagnosis by using clinical data,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.
- [164] T. Masquelier and S. J. Thorpe, “Unsupervised learning of visual features through spike timing dependent plasticity,” *PLOS Computational Biology*, vol. 3, pp. 1–11, Feb. 2007.
- [165] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

-
- [166] T. Kohonen, M. R. Schroeder, and T. S. Huang, eds., *Self-Organizing Maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 3rd ed., 2001.
- [167] T. Kohonen, “Essentials of the self-organizing map,” *Neural Networks*, vol. 37, pp. 52–65, 2013.
- [168] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE*, vol. 78, pp. 1464–1480, Sept. 1990.
- [169] J. C. Dunn, “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters,” *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.
- [170] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.
- [171] T. C. Cahoon, M. A. Sutton, and J. C. Bezdek, “Breast cancer detection using image processing techniques,” *IEEE International Conference on Fuzzy Systems*, vol. 2, pp. 973–976, 2000.
- [172] D.-R. Chen, R.-F. Chang, and Y.-L. Huang, “Breast cancer diagnosis using self-organizing map for sonography,” *Ultrasound in Medicine & Biology*, vol. 26, no. 3, pp. 405–411, 2000.
- [173] M. Markey, J. Lo, G. Tourassi, and C. Floyd, “Self-organizing map for cluster analysis of a breast cancer database,” *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 113–127, 2003.
- [174] H. M. Moftah, A. T. Azar, E. T. Al-Shammari, N. I. Ghali, A. E. Hassanien, and M. Shoman, “Adaptive k-means clustering algorithm for mr breast image segmentation,” *Neural Comput. Appl.*, vol. 24, pp. 1917–1928, June 2014.

- [175] S. H. Lee, J. H. Kim, K. G. Kim, S. J. Park, and W. K. Moon, *K-Means Clustering and Classification of Kinetic Curves on Malignancy in Dynamic Breast MRI*, pp. 2536–2539. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- [176] S. Dalmiya, A. Dasgupta, and S. K. Datta, “Application of wavelet based k-means algorithm in mammogram segmentation,” *International Journal of Computer Applications*, vol. 52, pp. 15–19, 2012.
- [177] A. Elmoufidi, K. Fahssi, S. Andaloussi, and A. Sekkaki, *Detection of Regions of Interest in Mammograms by Using Local Binary Pattern and Dynamic K-Means Algorithm*, pp. 11–18. Orb Academic Publisher, 2014.
- [178] E. S. Samundeeswari, P. K. Saranya, and R. Manavalan, “Segmentation of breast ultrasound image using regularized k-means (rekm) clustering,” in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 1379–1383, March 2016.
- [179] K. Rezaee, “Designing an algorithm for cancerous tissue segmentation using adaptive k-means cluttering and discrete wavelet transform,” in *Journal of Biomedical Physics and Engineering*, pp. 93–104, 2013.
- [180] B. Chandra, S. Nath, and A. Malhotra, “Classification and clustering of breast cancer images,” in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 3843–3847, 2006.
- [181] A. Lashkari and M. Firouzmand, “Early breast cancer detection in thermogram images using adaboost classifier and fuzzy c-means clustering algorithm,” *Middle East Journal of Cancer*, vol. 7, no. 3, pp. 113–124, 2016.
- [182] T. W. Nattkemper, B. Arnrich, O. Lichte, W. Timm, A. Degenhard, L. Pointon, C. Hayes, and M. O. Leach, “Evaluation of radiological features for breast tumour

- classification in clinical screening with machine learning methods,” *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 129–139, 2005.
- [183] L. A. Salazar-Licea, J. C. Pedraza-Ortega, A. Pastrana-Palma, and M. A. Aceves-Fernandez, “Location of mammograms roi’s and reduction of false-positive,” *Computer Methods and Programs in Biomedicine*, vol. 143, pp. 97–111, 2017.
- [184] K. Marcomini, A. Carneiro, and H. Schiabel, “Application of artificial neural network models in segmentation and classification of nodules in breast ultrasound digital images,” in *International Journal of Biomedical Imaging*, pp. 3843–3847, 2016.
- [185] Z. İşcan, Z. Dokur, and T. Ölmez, *Improved incremental self-organizing map for the segmentation of ultrasound images*, pp. 293–302. Dordrecht: Springer Netherlands, 2007.
- [186] X. Zhu, “Semi-supervised learning literature survey,” Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.
- [187] C. Li and P. Yuen, “Semi-supervised learning in medical image database,” *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 2035, pp. 154–160, 2001.
- [188] J.-B. Li, Y. Yu, Z.-M. Yang, and L.-L. Tang, “Breast tissue image classification based on semi-supervised locality discriminant projection with kernels,” *Journal of Medical Systems*, vol. 36, no. 5, pp. 2779–2786, 2012.
- [189] M. Ngadi, A. Amine, and B. Nassih, “A robust approach for mammographic image classification using nsvc algorithm,” *ACM International Conference Proceeding Series*, vol. Part F126741, pp. 44–49, 2016.

- [190] F. R. Cordeiro, W. P. Santos, and A. G. Silva-Filho, "A semi-supervised fuzzy growcut algorithm to segment and classify regions of interest of mammographic images," *Expert Systems with Applications*, vol. 65, pp. 116–126, 2016.
- [191] F. Cordeiro, W. Santos, and A. Silva-Filho, "Analysis of supervised and semi-supervised growcut applied to segmentation of masses in mammography images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 5, no. 4, pp. 297–315, 2017.
- [192] Z. Nawel, A. Nabiha, D. Nilanjan, and S. Mokhtar, "Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 1, pp. 53–62, 2016.
- [193] N. Zemmam, N. Azizi, and M. Sellami, "CAD system for classification of mammographic abnormalities using transductive semi supervised learning algorithm and heterogeneous features," *12th International Symposium on Programming and Systems, ISPS 2015*, pp. 245–253, 2015.
- [194] N. Zemmam, N. Azizi, N. Dey, and M. Sellami, "Adaptative s3vm semi supervised learning with features cooperation for breast cancer classification," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 4, pp. 957–967, 2016.
- [195] N. Zemmam, N. Azizi, and M. Sellami, "CAD system for classification of mammographic abnormalities using transductive semi supervised learning algorithm and heterogeneous features," in *2015 12th International Symposium on Programming and Systems (ISPS)*, pp. 1–9, April 2015.

- [196] M. Peikari, J. Zubovits, G. Clarke, and A. L. Martel, *Clustering Analysis for Semi-supervised Learning Improves Classification Performance of Digital Pathology*, pp. 263–270. Cham: Springer International Publishing, 2015.
- [197] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure,” *SIGMOD Rec.*, vol. 28, pp. 49–60, June 1999.
- [198] Y. Zhu, F. Li, T. J. Vadakkan, M. Zhang, J. Landua, W. Wei, J. Ma, M. E. Dickinson, J. M. Rosen, M. T. Lewis, M. Zhan, and S. T. C. Wong, “Three-dimensional vasculature reconstruction of tumour microenvironment via local clustering and classification,” *Interface Focus*, vol. 3, no. 4, 2013.
- [199] X. Liu, J. Shi, S. Zhou, and M. Lu, “An iterated laplacian based semi-supervised dimensionality reduction for classification of breast cancer on ultrasound images,” pp. 4679–4682, Aug. 2014.
- [200] M. A. Jaffar, “Deep learning based computer aided diagnosis system for breast mammograms,” in *International Journal of Advanced Computer Science and Applications (ijacsa)*, 2017.
- [201] M. G. Ertosun and D. L. Rubin, “Probabilistic visual search for masses within mammography images using deep learning,” in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1310–1315, Nov. 2015.
- [202] Y. Qiu, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng, “Computer-aided classification of mammographic masses using the deep learning technology: a preliminary study,” *Proc.SPIE*, pp. 9785–9791, 2016.
- [203] Y. Zheng, Z. Jiang, F. Xie, H. Zhang, Y. Ma, H. Shi, and Y. Zhao, “Feature extraction from histopathological images based on nucleus-guided convolutional neural

- network for breast lesion classification,” *Pattern Recognition*, vol. 71, pp. 14–25, 2017.
- [204] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” *PLOS ONE*, vol. 12, pp. 1–14, June 2017.
- [205] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A dataset for breast cancer histopathological image classification,” *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1455–1462, July 2016.
- [206] X. Jin and J. Han, *K-Means Clustering*, pp. 563–564. Boston, MA: Springer US, 2010.
- [207] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, pp. 790–799, Aug. 1995.
- [208] V. Vapnik, “Statistical learning theory, adaptive and learning systems for signal processing communications and control,” *John Wiley & Sons*.
- [209] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.
- [210] A. Graves, “Generating sequences with recurrent neural networks,” *CoRR*, 2013.
- [211] Y. Xiao and K. Cho, “Efficient character-level document classification by combining convolution and recurrent layers,” *CoRR*, vol. abs/1602.00367, 2016.
- [212] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, “Convolutional recurrent neural networks: Learning spatial dependencies for image representation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 18–26, June 2015.

- [213] H. Wu and S. Prasad, “Convolutional recurrent neural networks for hyperspectral data classification,” *Remote Sensing*, vol. 9, no. 3, pp. 1–20, 2017.
- [214] B. E. Bejnordi, G. C. A. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. J. S. Litjens, and J. van der Laak, “Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images,” *CoRR*, vol. abs/1705.03678, 2017.
- [215] <https://rdm.insectec.pt/dataset/nis-2017-003>.
- [216] K. Dimitropoulos, P. Barmpoutis, C. Zioga, A. Kamas, K. Patsiaoura, and N. Grammalidis, “Grading of invasive breast carcinoma through grassmannian vlad encoding,” *PLOS ONE*, vol. 12, pp. 1–18, Sept. 2017.
- [217] <http://www.who.int/mediacentre/factsheets/fs297/en/>
- [218] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, pp. 504–507, July 2006.
- [219] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012.
- [220] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [221] E. H. Land and J. J. McCann, “Lightness and retinex theory,” 1971.
- [222] D. J. Jobson, Z. Rahman, and G. A. Woodell, “A multiscale retinex for bridging the gap between color images and the human observation of scenes,” *IEEE Transactions on Image Processing*, vol. 6, pp. 965–976, Jul 1997.

-
- [223] M. Blot, M. Cord, and N. Thome, “Maxmin convolutional neural networks for image classification,” *CoRR*, 2016.
- [224] <http://web.inf.ufpr.br/vri/breast-cancer-database>
- [225] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, “A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images,” *Neurocomputing*, vol. 191, pp. 214–223, 2016.
- [226] H. Rezaeilouyeh, A. Mollahosseini, and M. H. Mahoor, “Microscopic medical image classification framework via deep learning and shearlet transform,” *Journal of Medical Imaging*, vol. 3, pp. 1–13, 2016.
- [227] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *CoRR*.
- [228] G. Strang, “The discrete cosine transform,” *SIAM Review*, vol. 41, pp. 135–147, 1999.
- [229] A. Bazzani, A. Bevilacqua, D. Bollini, R. Brancaccio, R. Campanini, N. Lanconelli, A. Riccardi, and D. Romani, “An svm classifier to separate false signals from microcalcifications in digital mammograms,” *Physics in Medicine and Biology*, vol. 46, no. 6, pp. 1651–1663, 2001.
- [230] G. Gatuha and T. Jiang, “Evaluating diagnostic performance of machine learning algorithms on breast cancer,” in *Revised Selected Papers, Part II, of the 5th International Conference on Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques - Volume 9243*, IScIDE 2015, (New York, NY, USA), pp. 258–266, Springer-Verlag New York, 2015.

-
- [231] S. Anand and R. A. V. Rathna, "Detection of architectural distortion in mammogram images using contourlet transform," pp. 177–180, March 2013.
- [232] F. Moayed, Z. Azimifar, R. Boostani, and S. Katebi, *Contourlet-Based Mammography Mass Classification*, pp. 923–934. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- [233] J. Jasmine, S. Baskaran, and A. Govardhan, "Nonsubsampled contourlet transform based classification of microcalcification in digital mammograms," *Procedia Engineering*, vol. 38, pp. 622–631, 2012.
- [234] F. Pak, H. R. Kanan, and A. Alihashi, "Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (nsct) and super resolution," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 2, pp. 89–107, 2015.
- [235] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on Image Processing*, vol. 14, pp. 2091–2106, Dec. 2005.
- [236] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, pp. 532–540, Apr. 1983.
- [237] G. Strang, "The discrete cosine transform," *SIAM Rev.*, vol. 41, pp. 135–147, Mar. 1999.
- [238] E. Brook, R. El-yaniv, E. Isler, R. Kimmel, R. Meir, and D. Peleg, "Breast cancer diagnosis from biopsy images using generic features and svms," *Technion—Israel Institute of Technology, Kesabsaba, Israel, Tech. Rep. CS-2008-07*, 2008.

- [239] B. Zhang, “Breast cancer diagnosis from biopsy images by serial fusion of random subspace ensembles,” in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 1, pp. 180–186, Oct. 2011.
- [240] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, *Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks*, pp. 411–418. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.
- [241] H. Wang, A. C. Roa, A. N. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, “Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features,” *Journal of Medical Imaging*, vol. 1, pp. 1–8, 2014.
- [242] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, “Breast cancer multi-classification from histopathological images with structured deep learning model,” in *Scientific Reports*, 2017.
- [243] P. Wang, X. Hu, Y. Li, Q. Liu, and X. Zhu, “Automatic cell nuclei segmentation and classification of breast cancer histopathology images,” *Signal Processing*, vol. 122, no. Supplement C, pp. 1–13, 2016.
- [244] K. Rajesh, S. Rajeev, and S. Srivastava, “Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features,” vol. 2015, pp. 1–14, 2015.
- [245] H. Su, F. Liu, Y. Xie, F. Xing, S. Meyyappan, and L. Yang, “Region segmentation in histopathological breast cancer images using deep convolutional neural network,” in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 55–58, April 2015.

- [246] K. Sirinukunwattana, S. Raza, Y.-W. Tsang, D. Snead, I. Cree, and N. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.
- [247] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, July 2006.
- [248] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Syst. Appl.*, vol. 46, pp. 139–144, March 2016.
- [249] J. Gao, Y. Guo, and M. Yin, "Restricted boltzmann machine approach to couple dictionary training for image super-resolution," in *2013 IEEE International Conference on Image Processing*, pp. 499–503, Sept. 2013.
- [250] M. A. U. Khan, T. A. Soomro, T. M. Khan, D. G. Bailey, J. Gao, and N. Mir, "Automatic retinal vessel extraction algorithm based on contrast-sensitive schemes," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–5, Nov. 2016.
- [251] T. M. Khan, D. G. Bailey, M. A. U. Khan, and Y. Kong, "Efficient hardware implementation strategy for local normalization of fingerprint images," *Journal of Real-Time Image Processing*, pp. 1–13, 2016.
- [252] P. Howarth and S. Rüger, *Evaluation of Texture Features for Content-Based Image Retrieval*, pp. 326–334. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.
- [253] F. Bianconi, A. Álvarez Larrán, and A. Fernández, "Discrimination between tumour epithelium and stroma via perception-based features," *Neurocomputing*, vol. 154, pp. 119–126, 2015.

- [254] Z. Wang, H. Liu, Y. Qian, and T. Xu, "Crowd density estimation based on local binary pattern co-occurrence matrix," in *2012 IEEE International Conference on Multimedia and Expo Workshops*, pp. 372–377, July 2012.
- [255] S. H. Wu, K. P. Lin, H. H. Chien, C. M. Chen, and M. S. Chen, "On generalizable low false-positive learning using asymmetric support vector machines," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 1083–1096, May 2013.
- [256] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1–14, 2015.
- [257] K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast cancer detection in digital mammograms," in *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, Sept. 2015.
- [258] J. Zhou, C. Feng, X. Liu, and J. Tang, "A texture features based medical image retrieval system for breast cancer," in *2012 7th International Conference on Computing and Convergence Technology (ICCT)*, pp. 1010–1015, Dec 2012.
- [259] P. Král and L. Lenc, "LBP features for breast cancer detection," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2643–2647, Sept. 2016.
- [260] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," *Machine Intelligence and Pattern Recognition*, vol. 16, pp. 403–413, 1994.