

Shearlet-based Descriptors and Deep Learning Approaches for Medical Image Classification

by

Sadiq Al-Insaif

A thesis submitted in partial fulfillment of the requirements for the
Doctorate in Philosophy degree in Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Sadiq Al-Insaif, Ottawa, Canada, 2021

Declaration

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of University of Ottawa's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

Abstract

In this Ph.D. thesis, we develop effective techniques for medical image classification, particularly, for histopathological and magnetic resonance images (MRI). Our techniques are capable of handling the high variability in the content of such images. Handcrafted techniques based on texture analysis are used for the classification task. We also use deep learning models but training such models from scratch can be a challenging process, instead, we employ deep features and transfer learning.

First, we propose a combined texture-based feature representation that is computed in the complex shearlet domain for histopathological image classification. With complex coefficients, we examine both the magnitude and relative phase of shearlets to form the feature space. Our proposed techniques are successful for histopathological image classification. Furthermore, we investigate their ability to generalize to MRI datasets that present an additional challenge, namely high dimensionality. An MRI sample consists of a large number of slices. Our proposed shearlet-based feature representation for histopathological images cannot be used without adjustment. Therefore, we consider the 3D shearlet transform given the volumetric nature of MRI data. An advantage of the 3D shearlet transform is that it takes into consideration adjacent slices of MRI data.

Secondly, we study the classification of histopathological images using pre-trained deep learning models. A pre-trained deep learning model can act as a starting point for datasets with a limited number of samples. Therefore, we used various models either as unsupervised feature extractors, or weight initializers to classify histopathological images. When it comes to MRI samples, fine-tuning a deep learning model is not straightforward. Pre-trained models are trained on RGB images which have a channel size of 3, but an MRI sample has a larger number of slices. Fine-tuning a convolutional neural network (CNN) requires adjusting a model to work with MRI data. We fine-tune pre-trained models and then use them as feature extractors. Thereafter, we demonstrate the effectiveness of fine-tuned deep features with classical machine learning (ML) classifiers, namely a support vector machine and a decision tree bagger. Furthermore, instead of using a classical ML classifier for the MRI sample, we built a custom CNN that takes both the 3D shearlet descriptors and deep features as an input. This custom network processes our feature representation end-to-end and then classifies an MRI sample. Our custom CNN is more effective in comparison to a classical ML on a hidden MRI dataset. It is an indication that our CNN model is less susceptible to over-fitting.

Acknowledgments

With gratitude and appreciation, I would like to thank my supervisor professor Jochen Lang. Thanks very much for your help and guidance throughout this research.

I would also like to thank professor Herna Viktor, professor Hussein Al Osman, professor Richard Dansereau, and professor Hassan Rivaz for examining my thesis and providing me with valuable feedback.

Thanks to my friends, and colleagues from VIVA lab.

Finally, thanks to my parents, siblings, nephews, and nieces. You have provided me great support and encouragement throughout the years of my study.

Table of Contents

List of Tables	viii
List of Figures	x
List of Abbreviations	xii
1 Introduction	1
1.1 Overview	1
1.1.1 Pathology Image Acquisition	1
1.1.2 Radiology Image Acquisition	2
1.1.3 Machine Learning in Medical Image Classification	2
1.2 Problem Statement	2
1.2.1 Traditional Learning-Based Techniques	3
1.2.2 Deep Learning-Based Techniques	4
1.3 Thesis Statement	6
1.4 Contributions	6
1.4.1 Medical Image Classification: Comparative Literature Review	6
1.4.2 Shearlet-based Techniques for Histological Image Classification . . .	6
1.4.3 Histological Image Classification using Deep Features and Transfer Learning	7
1.4.4 Classification of Alzheimer’s Disease based on MRI Data	7
1.5 Outline	8
2 Medical Image Classification: Comparative Literature Review	9
2.1 Overview	9
2.2 Histopathological Images	10

2.2.1	Histopathological Datasets	10
2.3	Magnetic Resonance Imaging (MRI) Data	15
2.4	Disease Association using Machine Learning	16
2.4.1	Tissue Classification	19
2.5	Early Diagnosis of Alzheimer’s Disease from MRI	24
2.6	Summary	27
3	Shearlet-based Techniques for Histological Image Classification	29
3.1	Overview	29
3.2	Related Work	30
3.3	Classification Algorithms	31
3.3.1	Support Vector Machine (SVM)	31
3.3.2	Decision Tree Bagger (DTB)	32
3.4	Method	32
3.4.1	Attributes from the Complex Shearlet Transform	33
3.4.2	Fusion of Feature Sets	37
3.4.3	Principal Component Analysis (PCA) for Feature Reduction	37
3.5	Experiments	37
3.5.1	Datasets	37
3.5.2	Cross-validation (CV)	38
3.6	Results and Analysis	38
3.7	Summary	47
4	Histological Image Classification using Deep Features and Transfer Learning	48
4.1	Overview	48
4.2	Deep Models For Histopathological Images	49
4.3	CNN as Feature Extractor	50
4.3.1	Infinite Latent Feature Selection	50
4.4	Transfer Learning For Histopathological Images	51
4.4.1	CNN Fine-tuning	51
4.5	Experiments	52
4.5.1	Datasets	52

4.5.2	Cross-validation (CV)	52
4.6	Results and Analysis	52
4.6.1	Classification Results Using Kather's dataset	53
4.6.2	Classification Results Using Epistroma, Warwick-QU, and BreakHis datasets	54
4.6.3	State-of-the-art Results	55
4.7	Summary	58
5	Classification of Alzheimer's Disease based on MRI Data	59
5.1	Overview	59
5.2	Proposed System and Components Configurations	61
5.2.1	Shearlet-based Descriptors for AD Classification	61
5.2.2	Fine-tuning Pre-trained CNN Models for AD Classification	64
5.2.3	Custom CNN	67
5.3	Experiments	68
5.3.1	Datasets	68
5.3.2	Nested Cross-validation (CV)	70
5.4	Results and Analysis	71
5.4.1	Classification Results Using Shearlet-Based Descriptors	71
5.4.2	Classification Results using Fine-tuned Pre-trained CNNs	73
5.4.3	Classification Results While Combining Shearlet-based Descriptors with Deep Features	74
5.4.4	State-of-the-art Results	77
5.5	Summary	80
6	Conclusion	82
6.1	Summary	83
6.2	Limitations and Future Work	85
	APPENDICES	87
	A Supplemental Material For Chapter 2	88
	B Supplemental Material For Chapter 3	99
	References	110

List of Tables

2.1	Summary of histopathological datasets (DB).	16
3.1	Summary of datasets used in Ch. 3.	37
4.1	Summary of datasets used in Ch. 4.	52
4.2	SVM classification performance utilizing deep features of various pre-trained models.	53
4.3	SVM classification performance utilizing top ranked deep features of various pre-trained models.	54
4.4	Classification performance utilizing pre-trained models as weight initialization, CNNs are fully fine-tuned.	56
4.5	Block-wise fine-tuning of pre-trained SqueezeNet for histopathological image classification.	57
4.6	Block-wise fine-tuning of pre-trained ResNet-18 for histopathological image classification.	58
5.1	ADNI classification performance using shearlet-based descriptors for (CN + sMCI) vs. (pMCI + AD).	71
5.2	OASIS classification performance using shearlet-based descriptors for CN vs. AD.	72
5.3	ADNI classification performance using deep features for (CN + sMCI) vs. (pMCI + AD).	73
5.4	OASIS classification performance using deep features for CN vs. AD.	74
5.5	ADNI classification performance while combining shearlet-based descriptors with deep features for (CN + sMCI) vs. (pMCI + AD).	75
5.6	Using unseen data samples of ADNI to test the generalization of our proposed techniques.	75
5.7	OASIS classification performance while combining shearlet-based descriptors with deep features for CN vs. AD.	76

5.8	Using unseen data samples of OASIS to test the generalization of our proposed techniques.	76
5.9	Comparative analysis of related work for AD classification.	79
5.9	Comparative analysis of related work for AD classification.	80
A.1	Summary of methods using BreakHis, Warwick-QU, USCB, Epistroma, or Kather (tumor & strom tissue only).	88
A.2	Summary of methods using multi-class Kather’s dataset.	92
A.3	Work summary of methods using BACH dataset.	93
A.4	Summary of methods using HistologyDS dataset.	94
A.5	Summary of methods using Lymphoma dataset.	96
A.6	Summary of methods using Path960 dataset.	97
B.1	Classification performance on Kather’s dataset using shearlet-based descriptors.	99
B.1	Classification performance on Kather’s dataset using shearlet-based descriptors.	100
B.1	Classification performance on Kather’s dataset using shearlet-based descriptors.	101
B.2	Classification performance on BreakHis dataset using shearlet-based descriptors.	102
B.2	Classification performance on BreakHis dataset using shearlet-based descriptors.	103
B.2	Classification performance on BreakHis dataset using shearlet-based descriptors.	104
B.3	Classification performance on Epistroma dataset using shearlet-based descriptors.	104
B.3	Classification performance on Epistroma dataset using shearlet-based descriptors.	105
B.3	Classification performance on Epistroma dataset using shearlet-based descriptors.	106
B.3	Classification performance on Epistroma dataset using shearlet-based descriptors.	107
B.4	Classification performance on Warwick-QU dataset using shearlet-based descriptors.	107
B.4	Classification performance on Warwick-QU dataset using shearlet-based descriptors.	108
B.4	Classification performance on Warwick-QU dataset using shearlet-based descriptors.	109

List of Figures

1.1	Traditional ML vs. DL for medical image classification.	3
2.1	Samples from the Epistroma dataset. One Example of stroma tissue (left) and one from epithelium tissue (right).	11
2.2	Samples from the Warwick-QU dataset. One example of benign tissue (left) and one from malignant tissue (right).	11
2.3	Samples from the BreakHis dataset. Top row shows benign samples and bottom row shows malignant samples. Columns represent magnification factors of $40\times$, $100\times$, $200\times$, and $400\times$, respectively.	11
2.4	Samples from the Kather's dataset. Top row shows samples of: tumour epithelium, simple stroma, complex stroma, lymphocytes, respectively. Bottom row shows samples of: debris, mucosal glands, adipose tissue, and background, respectively.	12
2.5	Samples from the Grand-Challenge Breast Cancer Images. Left to Right: normal, benign, in situ, and invasive, respectively.	12
2.6	Samples from the Lymphoma dataset. From left to right: CLL, FL, and MCL.	13
2.7	Samples from the histologyDS dataset. Left to right: connective, epithelial, muscular, and nervous tissue, respectively.	13
2.8	Samples from the CFTO dataset. Top row: Heart, Muscular, and Elastic respectively. Bottom row: Vein, Connective and Light, respectively.	14
2.9	Samples from the UCSB dataset. Benign (left) and Malignant (right). Each image in this dataset has a size of 896×768	15
2.10	Samples from the KIMIA PATH960 dataset. Each image in this dataset has a size of 308×168	15
3.1	Overview of our shearlet-based system for histopathological image classification. © Copyright IEEE 2019.	33
3.2	Classification performance on Kather's dataset using shearlet-based descriptors.	40
3.3	Classification performance on BreakHis dataset using shearlet-based descriptors.	41

3.4	Classification performance on Epistroma dataset using shearlet-based descriptors.	42
3.5	Classification performance on Warwick-QU dataset using shearlet-based descriptors.	43
3.6	Classification performance of SVM with an increment of 50 principal components (PCs) of best Fusion till all PCs are used.	44
3.7	Confusion matrices of our best performing proposed methods while using reduced set of features (see Figure 3.6).	45
3.8	Classification performance of SVM with an increment of 50 PCs of Fusion #3 till all PCs are used.	46
4.1	Our approach for using deep features for histopathological image classification. © Copyright IEEE 2020.	50
4.2	Varying the number of images per class of multi-class Kather's dataset (8-class dataset). Starting with 25 images per class till all available images are used. © Copyright IEEE 2020.	55
5.1	Our proposed system for integrating the 3D shearlet-based textural descriptors with deep features for the ADNI and OASIS MRI classification.	61
5.2	Descriptors extracted from one sample of the OASIS dataset. The descriptors are extracted again after the sample is rotated by 90°.	63
5.3	Our custom CNN model which is used to transform the combined shearlet-based descriptors with deep features.	67
5.4	Examples of MRI slices before (top) and after (bottom) processing. Left to right: CN, sMCI, AD, and pMCI, respectively.	68
5.5	Examples of the OASIS MRI slices. CN on the left and AD on the right.	69

List of Abbreviations

Acronym	Description	Acronym	Description
DT	Decision Tree [172]	RF	Random Forest [27]
PL	Polynomial Classifier [128]	SURF	Speeded up Robust Features [17]
LR	Logistic Regression [86]	NN	Nearest Neighbor [43]
CLBP	Complete Local Binary Pattern [75]	RBF	Radial Basis Function [164]
LPQ	Local Phase Quantization [158]	LTP	Local Ternary Patterns [211]
RICLBP	Rotation Invariant Co-occurrence Among Adjacent LBP [156]	CLAHE	Contrast Limited Adaptive Histogram Equalisation [168]
GF	Gabor Filters [54]	FV	Fisher Vector
LPQ-TOP	LPQ Three Orthogonal Planes	GLRLM	Gray-Level Run-Length Matrix [212]
GLDM	Grey-Level Difference Matrix [41]	LQP	Local Quinary Patterns
CENTRIST	Census Transform Histogram [229]	RBM	Restricted Boltzmann Machine
IOCLBP	Improved Opponent Color LBP [22]	LCC	Local Color Contrast [45]
ORB	Oriented FAST and Rotated BRIEF key points [182]	AHP	Adaptive Hybrid Pattern [243]
BSIF	Binarized Statistical Image Feature [102]	SIFT	Scale-Invariant Feature Transform [133, 127]
SFTA	Segmentation-based Fractal Texture Analysis [42]	LOSIB	Local Oriented Statistics Information Booster [67]
XGBoost	Extreme Gradient Boosting [35]	GBT	Gradient Boosting Trees [221]

Chapter 1

Introduction

1.1 Overview

Medical image analysis is time-consuming [239, 204]. Most importantly, different interpretation from different medical experts over the interpretation of images is not uncommon; therefore, computer-assisted diagnosis (CAD) can be a great asset to medical experts to help in making an objective diagnosis.

In particular, analyzing difficult cases of a pathological condition is crucial. Therefore, it is very beneficial to utilize the power of image processing coupled with machine learning/deep learning to pre-screen and identify normal from abnormal cases. Our main goal of this thesis is to build CAD systems that are effective and efficient to study medical images in pathology and in radiology which are integral for many cancer diagnoses in the field of medicine [203].

1.1.1 Pathology Image Acquisition

Pathology is the study of disease. More particularly, histopathology is the study of cells and tissues under the microscope. Such diagnostic is considered a gold standard for the detection of most lethal diseases [180]. A histopathologist starts with a tissue biopsy from an organism/organ [84]. A biopsy is a procedure during which tissue is removed. The removed tissue from an organ is placed in a fixative prior to examination under the microscope. To be able to differentiate between different cellular levels under the microscope, the tissue is stained. Typically, the tissue is stained with hematoxylin and eosin (H & E). A pathologist then examines the histological slide to report their findings according to their expertise.

In order to be able to apply computational techniques, the histological slides first need to be digitized [62, 225]. As such, the digitization of histological slides can be either microscopy images (i.e., contain one pathological condition) or whole slide image (WSI) (i.e., typically contains multiple pathological regions). Therefore, WSI is typically annotated on a pixel level for segmentation purposes. However, WSI can be tiled into patches (e.g.,

256×256), and then a patch can be given a label as a whole based on the dominant pathological condition.

1.1.2 Radiology Image Acquisition

Radiology (or so-called diagnostic imaging) requires capturing images utilizing various modalities. Particularly, magnetic resonance imaging (MRI) [227] can be used by a radiologist to obtain images of the structure of an organ. To capture such anatomy, MRI scanners use strong magnets that generate a powerful magnetic field, in which the protons in the human body are forced to orient with [20]. When the magnetic field is turned off, the MRI sensors determine the time needed for the protons to realign to their original position. Essentially, MRI captures the fluid characteristics of various human tissues, which are in turn computer-processed to generate black and white images. Thereafter, a medical expert in the field of radiology examines various tissues based on MRI images to make a diagnosis.

Unlike histopathology, where molecular and cellular components of tissue are captured under the microscope, MRI scanners are digital in nature [20]. Furthermore, MRI scans of a patient can yield hundreds of slices as opposed to a histological slide which is typically a single image, but with high resolution. Intuitively, with MRI images, we must not only account for the wide range of variability across different slices, but we also need to account for the computational complexity to deal with such a large number of slices.

1.1.3 Machine Learning in Medical Image Classification

Having medical images in digital form allows for building CAD applications for image processing/analysis techniques [92, 18, 39]. A common practice for image association between different medical images is the use of machine learning (ML), which can be categorized as supervised or unsupervised ML. In supervised learning, the training process consists of learning the distribution of an image x given that it belongs to class y , in other words, we train a model on a set of images where the ground truth label is known. Several well-known supervised learning algorithms are used extensively in the literature for medical image classification: multilayer perceptron (MLP) [80], support vector machine (SVM) [218, 169], and others. On the other hand, in the absence of ground truth labels, unsupervised learning can be adopted where a mapping function is devised to learn the hidden patterns, for example, k-means clustering [111] or principal component analysis (PCA) [100].

1.2 Problem Statement

The focus of this thesis is to construct an effective medical image classification model. In the context of traditional learning-based approaches, the robustness of a classifier is largely determined by the feature representation. In the literature, various methods have been proposed, for example, local phase quantization, scale-invariant feature transform,

speeded up robust features, and others. Techniques that utilize signal processing typically choose to transform an image using the Fourier transform or a wavelet transform. Then, various attributes are computed from the transform’s coefficients. Nevertheless, such transforms have limitations. The Fourier transform does not incorporate the spatial information of a transformed image. Also, the wavelet transform lacks directional sensitivity. Therefore, our traditional descriptors are based on the shearlet transform [115]. The shearlet transform addresses the aforementioned limitations and provides a sparse representation of an image at multiple scales and multiple directionalities. As a result, it is an effective choice for studying medical images, for instance, histopathological images to study the intra-relationship of cells. Additionally, the complex coefficients of shearlets are used to obtain both the magnitude and phase coefficients. Such coefficients are summarized utilizing various texture-based methods to obtain distinctive descriptors. For instance, the co-occurrence matrix, which is a 2D histogram, is computed for the directional sub-bands, and potentially leading to training an effective classification model.

When it comes to deep learning (DL), the effectiveness of a DL model depends on the architecture of a model. However, training a supervised DL model requires an abundance of samples from each class, especially for deeper models. However, there is a lack of well-annotated medical images, which makes the task of training a DL model from scratch challenging. Data augmentation helps to generate enough training samples, but it is not clear whether applying data augmentation helps predict unseen data samples. Therefore, we instead fine-tune deep learning models without applying data augmentation.

1.2.1 Traditional Learning-Based Techniques

An overview of the generic approach using a traditional learning-based technique is outlined in Figure 1.1(a). In traditional learning-based methods, the image needs typically some pre-processing. Next, features need to be extracted from the medical images. Then,

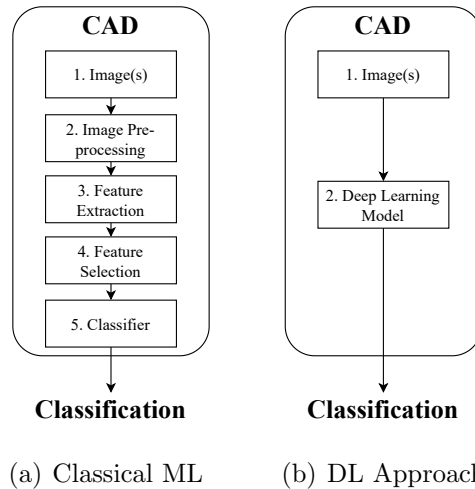


Figure 1.1: Traditional ML vs. DL for medical image classification.

these extracted features are used for training and building a model to discriminate between different types of abnormalities. Many studies have proposed different techniques to study the contents of digitized histological slides and MRI samples. A typical approach is a texture-based analysis, for example, local binary pattern (LBP) [157], gray-level co-occurrence matrix (GLCM), and many other techniques [92, 18]. As such, analytic and logical algorithms are utilized to study the content of medical images [126, 103, 204]. In the field of texture analysis, techniques are devised to examine medical image content through the computation of local or global statistical information to reveal repeated patterns, e.g., size or color [156, 67]. Commonly, that statistical information is computed on the individual pixels in the spatial domain.

Given the complex content of medical images, we model the complexity of histopathological tissue prediction by proposing a novel hand-crafted feature representation based on the shearlet transform [115]. We summarize the shearlet coefficients using textural methods, namely, co-occurrence texture feature [198, 77], LBP [157], local oriented statistics information booster (LOSIB) [67], and segmentation-based fractal texture analysis (SFTA) [42]. In our research, we investigate both parametric and non-parametric classifiers: SVM or decision tree bagger (DTB) to distinguish the tissue types. Our method calculates various local and global attributes that reveal complex patterns in the histopathological tissues. We successfully demonstrate that the combination of some sets of features is capable of generating a robust feature space. Thereafter, we employ principal component analysis to further improve the classification performance while having a reduced set of features.

We further examine the possibility to extend our shearlet-based techniques to analyze MRI samples. In our application of MRI data classification, we apply our techniques to the detection of Alzheimer’s disease. However, MRI samples present a different challenge because of the high number of slices. Therefore, given the volumetric nature of MRI data, we propose to use the 3D shearlet transform. However, before applying the 3D shearlet transform, each MRI sample is pre-processed with the following steps: 1) skull stripping, 2) image normalization, and 3) cropping. Then, we transform the MRI sample to the 3D shearlet domain in which we extract various texture descriptors, i.e., the same four texture methods we used to study histopathological images.

1.2.2 Deep Learning-Based Techniques

Recently, DL techniques [188, 15, 174, 18, 181] have been developed for the classification of medical images. DL uses a sequence of operations that are generic and can be applied directly to any type of raw input images, including medical images. A typical class of DL is a CNN which can be used for image classification [121]. DL models are capable of learning and discovering multiple levels of representations. As such, lower-level layers typically encode feature maps of edges and color, and higher-level layers learn semantic representations of the data [189, 234]. For instance, pixels arrangement comes in the form of an array that can be fed to a DL model. The presence or absence of edges at particular directions and positions in the image are learned in the first layer [121]. The second layer

would capture the motifs in the image content by detecting certain arrangements of edges. As a result, the deeper layers would capture objects as a mixture of these representations, i.e., higher and more abstract representation. Primarily, DL can learn these representations without any engineering, rather captured from data. With the advent of new learning algorithms, CNN architectures, and the availability of computational resources, DL models will reveal more successes in the field of image classification.

Therefore, many studies research the applicability of transferring knowledge to the domain of medical imaging [240, 192]. Alternatively, a pre-trained CNN model can be used (without any fine-tuning) as an unsupervised feature extractor [174]. Furthermore, some studies combine hand-crafted and deep features extracted from pre-trained models [149].

However, the progress of deep learning techniques in medical image classification is hampered by the scarcity of samples of bio-medical images [171]. This fact imposes a challenge for training deep models from scratch as abundant samples are a must to achieve state-of-art results [193]. State-of-the-art CNNs, e.g., DenseNet [89], have achieved high accuracy when trained and tested on ImageNet [46] but ImageNet contains more than a million images. Nevertheless, deep features and transfer learning both have enabled the use of these deep models in different domains including medial applications [91, 4]. Hence, we have investigated these techniques for histopathological images. We investigate the use of pre-trained models [93, 186] as unsupervised feature extractors [149]. We also investigate pre-trained models as weight initializers either by fine-tuning weights of all layers or freezing some layers to reduce training time [240] and to avoid a model memorizing the training samples (i.e., over-fitting) and fail to generalize on new emerging histopathological images. Both of our hand-crafted and deep learning techniques have shown excellent classification results when examined on four histopathological datasets that present different levels of complexity.

Similar to our traditional shearlet-based approach, we have extended our descriptors to examine their ability to generalize for radiology, namely to MRI image classification for Alzheimer’s disease. It is worth noting that because of the high number of slices per MRI sample and the limited number of samples, deep learning models will suffer from high computational cost and lack of generalization. Further, using unsupervised deep learning to obtain a feature vector per slice will yield a high-dimensional feature representation for an MRI sample. To solve these problems, we fine-tune a deep learning model and then explore such a model for its capabilities for generating descriptors that encode characteristics to distinguish Alzheimer’s disease from cognitively normal patients. In other words, we leverage the capabilities of a CNN model to learn feature maps from stacked slices of an MRI sample to create the feature space.

Furthermore, in an attempt to provide a more effective MRI sample classification model, we propose to investigate the integration of both the 3D shearlet descriptors and deep features. After concatenating features vectors that are computed from the shearlet domain and from a CNN model, we then propose to train a classification model in the following manner: 1) training a classical classifier (i.e., SVM and DTB); 2) building a custom CNN model to post-process feature vectors end-to-end.

1.3 Thesis Statement

Traditional machine learning and deep-learning techniques have been proposed to classify medical images in the literature but the utility of these approaches remain often unclear. Therefore, our main goal in this thesis is to develop, apply, and comparatively evaluate techniques that are capable of effectively and efficiently classify medical images. The techniques developed in this thesis should be able to handle 1) the scarcity of available data for training and testing; 2) the high inter and intra content variability; and 3) the high-dimensionality of medical images. Above all, these techniques should be (with minor adjustments) usable for different medical image classification tasks in histopathology and in radiology. Our goal is to demonstrate that our techniques apply to classification tasks where classical hand-crafted texture-based features have shown promise in the literature.

1.4 Contributions

1.4.1 Medical Image Classification: Comparative Literature Review

The goal of this chapter is to provide a review of techniques that are tailored for medical image classification. Furthermore, a detailed description of both classical and deep learning techniques provides a basis for comparison of our proposed techniques with the state-of-the-art. In this regard, we have researched techniques that can be applied to both radiology and pathology datasets.

A detailed summary of those existing techniques to study medical images is given in Chapter 2. Part of this chapter is submitted for publication and still under review [9].

1.4.2 Shearlet-based Techniques for Histological Image Classification

We propose a technique taking advantage of the multi-directionality and multi-scale properties of the shearlet transform for automatic classification of different tissue types in histopathological images. Our approach is to combine different sets of features into a novel texture feature space based on the complex shearlet domain. We have not only examined the magnitude coefficients, but we also incorporate the relative phase (RP) coefficients to form the feature representation. Our shearlet feature space consists of co-occurrence texture features, Local Binary Patterns, Local Oriented Statistic Information Booster, and segmentation-based Fractal Texture Analysis. These descriptors are also used jointly to increase the classification accuracy of histopathological images when used with classical machine learning.

A detailed analysis of this contribution is available in Chapter 3. This study led to a publication in an IEEE conference and an extended journal publication [5, 7].

1.4.3 Histological Image Classification using Deep Features and Transfer Learning

Supervised deep learning techniques can not be applied to histological image classification without some adjustment due to the major challenge of having a very limited number of available well-annotated data. Therefore, we provide strategies of two techniques to deal with limited domain data: using deep features and fine-tuning a CNN model. We examine the following state-of-art CNN models: SqueezeNet-v1.1, MobileNet-v2, ResNet-18, and DenseNet-201.

We demonstrate that using feature vectors that are extracted from one of four CNN models with a classical SVM for training and testing can lead to higher accuracy than previously published results on the following publicly available datasets: Warwick-QU, Epistroma, BreakHis, multi-class Kather. Similar results can be obtained by fully fine-tuning the aforementioned CNN models. We also study the effectiveness of block-wise fine-tuning of two models (i.e., SqueezeNet-v1.1 and ResNet-18) and show that it is not necessary to fully fine-tune leading to savings in time and space.

Details about our work in deep learning are presented in Chapter 4. This work led to a publication in an IEEE conference [6].

1.4.4 Classification of Alzheimer’s Disease based on MRI Data

In this study, we rely on an MRI dataset. This work extends our techniques based on shearlet descriptors and deep learning strategies to classify the MRI dataset. Both of our hand-crafted and deep learning strategies require limited changes. Given the volumetric nature of MRI data, we use the 3D shearlet transform. Then, the shearlet coefficients are summarized using texture-based methods.

Further, we propose to leverage the capabilities of CNN’s model mathematical operations to learn feature maps from stacked slices of an MRI sample. Our starting point is pre-trained CNN models. A CNN model needs to be modified (particularly, a new input layer to fit the MRI data dimensions, and the convolution layer following the input layer). As such, we fine-tune SqueezeNet-v1.1, MobileNet-v2, Xception, and Inception-v3. Then, we explore such models for their capabilities for generating very compact descriptors of high-dimensional MRI data to classify Alzheimer’s disease samples from cognitively normal samples. Unlike if we were to use a DL model as an unsupervised feature extractor to compute a feature vector for each slice, and then combine all feature vectors to form the feature space of an MRI sample; this will generate a high-dimensional feature representation.

Finally, in an attempt to provide an effective classification model, we combine both shearlet-based descriptors and deep features. One way is simply concatenating the feature vectors and feed them to a classifier. Alternatively, after we concatenate the feature vectors are then fed to a custom CNN model in which the descriptors will be processed end-to-end to obtain the classification results.

Details about this work are presented in Chapter 5. This chapter is submitted for publication and still under review [8].

1.5 Outline

The rest of this Ph.D. thesis is organized as follows. Chapter 2 provides a review of techniques for medical image classification. Chapter 3 presents our proposed techniques for histopathological image classification in the complex shearlet domain. Chapter 4 presents our deep learning strategies for histopathological image classification. Chapter 5 extends our proposed techniques from histopathological images to MRI images. Finally, Chapter 6 concludes and summarizes our work.

Chapter 2

Medical Image Classification: Comparative Literature Review

2.1 Overview

In this chapter, we provide an in-depth review of existing techniques for medical image classification. To achieve a systematic review of techniques that are classifying histopathological images, we first consider a set of 10 histopathological datasets. These datasets are publicly available and are tailored for classification tasks. These datasets can be used as a benchmark for testing and evaluating new methodologies. We also present a review of publications that have utilized these datasets. The methods introduced in these papers are summarized and categorized as classical machine learning (CML) and deep learning (DL) approaches, and we report the accuracy of each method evaluated based on the benchmark datasets we have reviewed.

Many computational techniques that are applied to histopathological image classification can be extended to be used on other types of medical datasets, for example, MRI. Therefore, a selected number of publications that study early diagnosis of Alzheimer’s disease (AD) are reviewed toward the end of this chapter.

The organization of this chapter as follows: we first present our motivation and focus on the type of targeted publications in relation to histopathological image analysis as presented in Section 2.2. In this manner, we review publicly available histopathological datasets and provide access links (see Section 2.2.1). In Section 2.3, we provide a summary of two well-known data repositories that provide MRI scans of AD patients. These summarized datasets are tailored for classification tasks; therefore, in Section 2.4, we provide an overview of how generic a machine learning pipeline can be built for medical image classification. Then, we divide our review of medical image classification techniques into two categories: CML and DL. In Section 2.4.1, we summarize histopathological image classification publications for the reviewed publicly available datasets. For each summarized method, we provide the classification performance while using the corresponding benchmark dataset. As such, this will act as a baseline for our new proposed framework in

Chapter 3 and Chapter 4. Furthermore, recent publications that are investigating computational techniques related to AD that can act as our baseline for comparison purposes for Chapter 5 are included in Section 2.5.

2.2 Histopathological Images

The wide availability of histological slide scanners [3] makes histopathological images available in digital form for research. Image datasets enable the development of techniques for cancer identification, yet this remains a difficult process due to the complexity levels that are present in histopathological images [14].

The various types of tissue samples that are obtained from human tumors are complex. Manual examination of the tissue types to assess the progress of the disease is indispensable, but it is time-consuming and expensive. There is, therefore, a need to automate the analysis of histopathological images. Many studies have proposed identifying the tissue types from the pattern revealed from the hematoxylin and eosin (H & E) stained slides of the tissue. The various techniques can be categorized as: 1) traditional learning-based techniques (i.e., using CML and classical texture analysis); 2) DL (i.e., using deep features by fine-tuning or training models from scratch, or combining traditional descriptors with deep features), as we summarize in Section 2.4.1.

2.2.1 Histopathological Datasets

Reviews have been published previously [87, 23] that consider texture datasets from different areas including bio-medical textured datasets of different types. However, we provide a review of **histopathological datasets** that are curated and can be used for image-level classification tasks. The summary of each dataset reviewed in this section is reported in Table. 2.1.

Epistroma

Epistroma dataset contains tissue samples that are collected from a set of 643 patients with colorectal cancer who did their surgical operations at Helsinki University Central Hospital (Helsinki, Finland) [126]. This is a binary dataset (examples of each tissue type are shown in Figure 2.1, such that it contains tissues of variable sizes from stroma (551 samples) and epithelium (825 samples).

Warwick-QU Dataset

Warwick-QU is a binary dataset of colon histology sections. This dataset consists of tissues from benign (74 samples) and malignant (91 samples) [196]. Examples of both tissues are shown in Figure 2.2). Each tissue sample is obtained with a magnification factor of $20\times$.

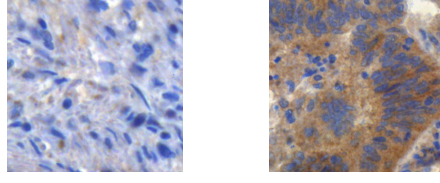


Figure 2.1: Samples from the Epistroma dataset. One Example of stroma tissue (left) and one from epithelium tissue (right).

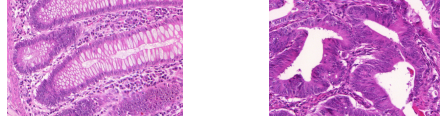


Figure 2.2: Samples from the Warwick-QU dataset. One example of benign tissue (left) and one from malignant tissue (right).

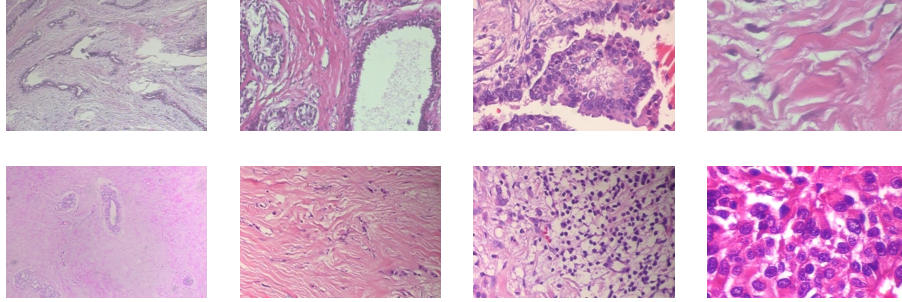


Figure 2.3: Samples from the BreakHis dataset. Top row shows benign samples and bottom row shows malignant samples. Columns represent magnification factors of $40\times$, $100\times$, $200\times$, and $400\times$, respectively.

Breast Cancer Histopathological Database (BreakHis)

BreakHis dataset contains a total of 7909 tissues primarily divided into benign and malignant breast tumor tissues [204]. Furthermore, each tissue type is divided into subcategories. The benign tissues have the following four subtypes: adenosis (A), fibro-adenoma (F), phyllodes tumor (PT), and tubular adenoma (TA). However, the malignant tissues have the following sub-types: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). These tissue types are obtained from breast tissue surgical open biopsy slides from January 2014 to December 2014. Each image has a size of 700×460 pixels, as a result of digitization with various magnification factors: $40\times$, $100\times$, $200\times$, and $400\times$. The distribution of benign/malignant samples for each magnification factor is as follows: 625/1370($40\times$), 644/1437($100\times$), 623/1390($200\times$) and 588/1232($400\times$) - collected from 82 patients. Examples of benign and malignant tissues of different magnification factors are shown in Figure 2.3.

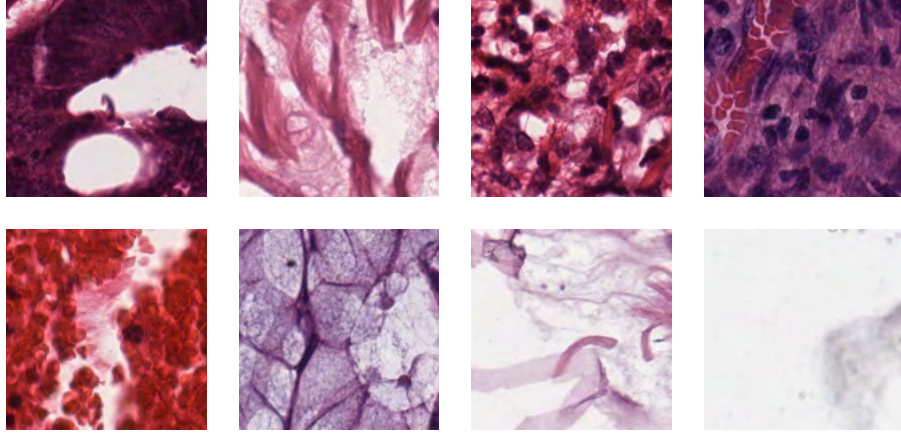


Figure 2.4: Samples from the Kather’s dataset. Top row shows samples of: tumour epithelium, simple stroma, complex stroma, lymphocytes, respectively. Bottom row shows samples of: debris, mucosal glands, adipose tissue, and background, respectively.

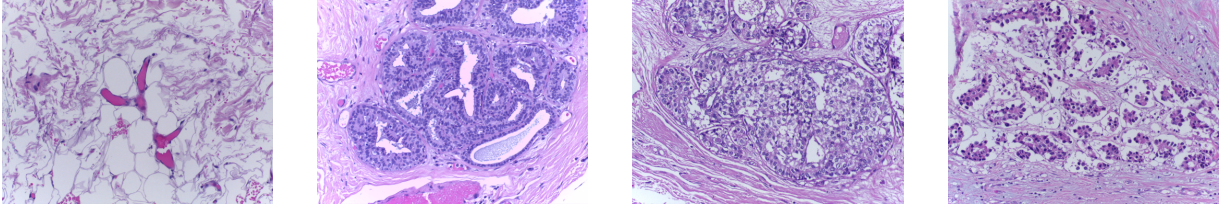


Figure 2.5: Samples from the Grand-Challenge Breast Cancer Images. Left to Right: normal, benign, in situ, and invasive, respectively.

Multi-class Kather’s

This dataset was collected by the University Medical Center of Mannheim (Heidelberg University, Mannheim, Germany) [103]. The dataset contains 5000 non-overlapping tissue images with a fixed size of 150×150 pixels. These images belong to eight different types of tissues, and each tissue type has 625 images. The eight types of tissues are tumor epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal glands, adipose tissue, and background (no tissue) as shown in Figure 2.4.

Grand-Challenge Breast Cancer Histology Images (BACH)

This dataset is a follow-up on a dataset released during the Bioimaging 2015 Grand Challenge (i.e, the 2015 dataset consists of 249 images). The new dataset, released in 2018, consists of 400 high-resolution breast cancer images (2048×1536 pixels) which can be requested from the organizers of Grand-Challenge BACH [14]. The collected images are categorized as follows: normal, benign, in situ carcinoma, or invasive carcinoma (100 samples for each category), collected between 2014 and 2017 from patients having breast cancer who are living in Portugal. Samples from each tissue type are shown in Figure 2.5.

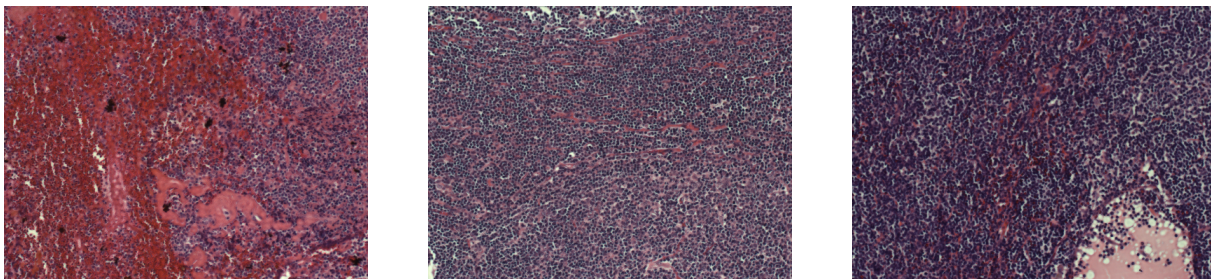


Figure 2.6: Samples from the Lymphoma dataset. From left to right: CLL, FL, and MCL.

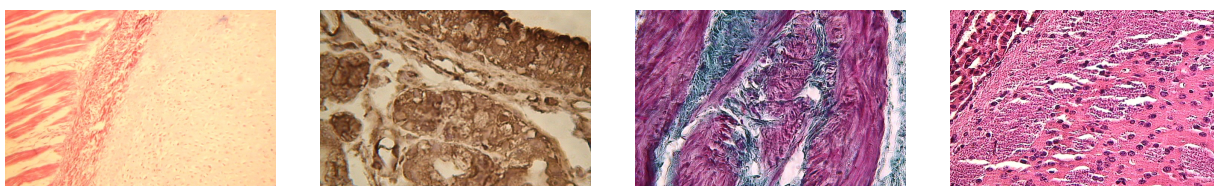


Figure 2.7: Samples from the histologyDS dataset. Left to right: connective, epithelial, muscular, and nervous tissue, respectively.

Image Informatics and Computational Biology Unit (IICBU) Repository

This repository contains a package of biological datasets [191]. These datasets are made publicly available for research and comparison purposes. Nevertheless, we are going to study only the textured dataset stained with H & E particularly to humans (i.e., Lymphoma) from this repository.

Lymphoma

The Lymphoma dataset consists of 375 images of lymph nodes (focused on lymph nodes cancer) and each image has a resolution of 1388×1040 pixels. This dataset consists of three classes: Chronic Lymphocytic Leukemia (CLL), Follicular Lymphoma (FL), and Mantle Cell Lymphoma (Figure 2.6).

The Histology Image Dataset (histologyDS)

The histologyDS [44] collected from different organs presents fundamental tissues of four types. This dataset consists of 2828 images with the following distribution for each class: 484 connectives, 804 epithelial, 514 muscular, and 1026 nervous tissue images as shown in Figure 2.7. These images are captured at different magnification factors of $10\times$, $20\times$, and $40\times$.

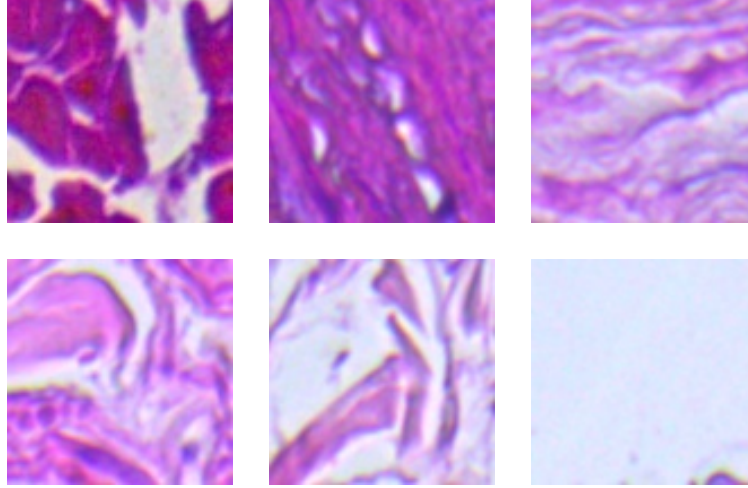


Figure 2.8: Samples from the CFTO dataset. Top row: Heart, Muscular, and Elastic respectively. Bottom row: Vein, Connective and Light, respectively.

Cardiovascular of Fundamental Tissues and Organs (CFTO)

Mazo et al. [139] published a dataset of histological cardiovascular tissues that consists of 6000 images divided into 70% for training and 30% for verification. This dataset belongs to 6 classes: (1) cardiac muscle (Heart), (2) muscular artery (Muscular), (3) elastic artery (Elastic) (4) muscle of the large vein (Vein), (5) connective tissue (Connective), (6) light regions (Light) as shown in Figure 2.8. There are 1000 images per class where each image has a resolution of 100×100 pixels. A separate set of 7500 images are provided for testing. Both sets of the different tissue types are obtained with a magnification factor of $10\times$.

UCSB Breast Cancer Cells

UCSB dataset contains samples of stained histopathology breast cancer cells [69], which is composed of two classes (benign and malignant). It was made publicly available by the center for bio-image informatics at the University of California, Santa Barbara (UCSB). The distribution of benign/malignant images in this dataset is 32/26 where each image has a resolution of 896×768 pixels (Figure 2.9).

KIMIA PATH960

In this dataset (Path960), the histopathological images (i.e., each image has a size of 308×168) are selected mainly based on visual clues that present various textures [114]. There are 48 histopathological images per class (i.e., there are 20 classes). As such, these images are extracted from more than 400 WSIs of muscle, epithelial, and connective tissues. Examples from each class are shown in Figure 2.10.

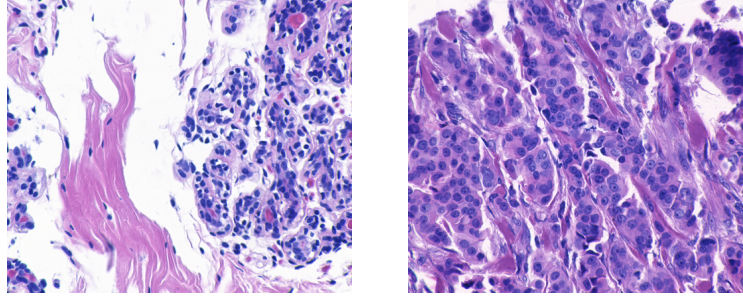


Figure 2.9: Samples from the UCSB dataset. Benign (left) and Malignant (right). Each image in this dataset has a size of 896×768 .

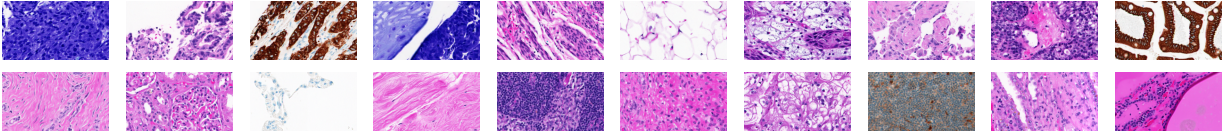


Figure 2.10: Samples from the KIMIA PATH960 dataset. Each image in this dataset has a size of 308×168 .

2.3 Magnetic Resonance Imaging (MRI) Data

Another significant application area of machine learning is MRI sample classification [135, 56]. Particularly, the early diagnosis of Alzheimer’s disease from MRI samples is critical for the proper course of treatment [214]. Many studies have been published exploring the classification of Alzheimer’s disease, i.e., utilizing publicly available datasets from Alzheimer’s Disease Neuroimaging Initiative (ADNI) or The Open Access Series of Imaging Studies (OASIS).

Many studies have utilized the ADNI and OASIS datasets for building a CAD system for the automatic classification of AD. In this section, we provide a brief description of both data repositories (i.e., the ADNI and OASIS).

Alzheimer’s Disease Neuroimaging Initiative (ADNI)

The ADNI data repository (adni.loni.usc.edu) is a multisite study for studying Alzheimer’s disease [96]. This data repository consists of MRI scans of three groups of subjects: 1) one group that has AD, 2) another group has a mild cognitive impairment (MCI), 3) and the final group is the cognitively normal (CN) participants. However, the MCI group can be further divided into two subgroups: 1) patients with a stable diagnosis of mild cognitive impairment (sMCI) that might remain stable MCI or revert to normal cognition, 2) patients who showed any sign of MCI diagnosis who might progress to AD (pMCI).

In this regard, the ADNI data repository is commonly used for the following binary classification tasks: 1) AD vs. CN, 2) pMCI vs. sMCI, 3) CN vs. MCI, 4) AD vs. MCI.

Table 2.1: Summary of histopathological datasets (DB).

DB	#Classes	#Samples	Sample Size	URL
Epistroma	2	1376	Variable	http://fimm.webmicroscope.net/Research/Supplements/epistroma/
Warwick-QU	2	165	Variable	https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/download/
BreakHis	2	7909	700× 460	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
Kather’s	8	5000	150× 150	https://zenodo.org/record/53169/
BACH18 (BACH15)	4	400 (249)	2048× 1536	https://iciar2018-challenge.grand-challenge.org/Dataset/
Lymphoma	3	375	1388× 1040	https://ome.grc.nia.nih.gov/iicbu2008/lymphoma/
HistologyDS4		2828	720× 480	http://online.unillanos.edu.co:8084/histologyDS/
CFTO	6	6000	100× 100	http://biscar.univallle.edu.co/?page_id=1003/
UCSB	2	58	896× 768	https://bioimage.ucsb.edu/research/bio-segmentation/
Path960	20	960	308× 168	https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path960/

The Open Access Series of Imaging Studies (OASIS)

The OASIS data repository (<https://www.oasis-brains.org/>) provides two types of data: cross-sectional and longitudinal MRI scans. However, our focus in this thesis is on cross-sectional MRI scans. This dataset was collected from 416 subjects whose ages range from 18 to 96.

The data repository provides a clinical dementia rating (CDR) score to determine the severity of dementia [137]. The CDR score ranges between 0 to 2. The assumption that the patient with a score of 0 implies a healthy subject. However, if the CDR is greater than 0, then it implies a patient with Alzheimer’s disease. In this regard, the OASIS data repository is commonly used for the following binary classification task: AD vs. CN.

2.4 Disease Association using Machine Learning

According to Michell [143] supervised machine learning relates the following main components: a task T, the experience E, and the performance P on the task. Hence, a hand-crafted or deep learning model for the task T maps any given histopathological tissue to

an output label. Such a model is built based on the experience E , i.e., a training dataset of histopathological tissues where each tissue type is labeled. ML algorithms are data-driven and optimized to generate an accurate classification (i.e., the performance P) using optimization algorithms. The ideal ML model is able to generalize when it is tested on unseen samples. Other types of learning techniques exist in the literature [170]. These techniques are especially beneficial for tackling the problem of small data classification.

- **Unsupervised learning:** such techniques are trained based on unlabeled datasets. As such, these techniques seek to obtain useful patterns from the data, i.e., a collection of examples that exhibit similar patterns are grouped together. According to Qi and Luo [170], some of the common unsupervised techniques are generative models (e.g., auto-encoders and generative adversarial nets (GANs), and self-supervised methods (e.g., autoregressive models and self-supervised representation learning either from images or videos). The generative models consist of two main components, i.e., the encoder that infers the representation from the data and the decoder reconstructs the data, but with sharper reconstruction. In the context of self-supervised learning techniques, they can be beneficial in predicting missing information from parts of the data while using the same shared latent representation.
- **Semi-supervised learning:** such techniques exploit datasets with both labeled and unlabeled examples to train a model. Qi and Luo [170] reviewed this category from two perspectives. Semi-supervised generative models, which is similar to unsupervised generative models, but here both labeled and unlabeled examples are used. However, the authors state that teacher-student models achieve state-of-the-art results. As such, a single or ensemble of teacher models are trained while using unlabeled samples. Then the predicted labels (or pseudo labels) are utilized to obtain a supervised student model. It is necessary to ensure the consistency between the predictions of teacher and student.

A commonly adopted evaluation technique (either evaluating CML or DL) is to use cross-validation (CV) to obtain robust statistical results. CV measures the performance of a classifier system by splitting the data into training and testing, such that the testing data is hidden during the training. A dataset is split into K folds, where $K - 1$ folds are used for training, and 1-fold is used for testing. The training is repeated K times with different testing folds. Thus, to evaluate the performances of the different classifiers, the following common performance metrics of a classification model in each fold are computed: accuracy, sensitivity, specificity, and area under receiver operating characteristic curve (AUC), i.e., a plot of true positive rates versus false-positive rates.

To build a robust classification model in the case of CML, the extraction of useful knowledge from an image is a critical step. While the above applies to both histopathological images and MRI, in the remainder of this section we describe the specifics of the classification of histopathological images. Examples of common attributes are color, morphological, or textural descriptors [124]. Consequently, those extracted feature vectors are used to classify the histopathological images. On the other hand, the main advantage of deep learning approaches is their ability to learn features from raw input histopathological

images end-to-end, passing the feature extraction step. In essence, to build a CAD system to distinguish normal from abnormal samples using machine learning, one needs the following steps:

- Step 1: Histopathological datasets are acquired.
- Step 2: Pre-processing: denoising, color deconvolution, color format conversion, stain normalization.
- Step 3: Feature extraction and feature selection.
- Step 4: Applying classification.

Step 1 is complete for publicly available datasets and enables practitioners to focus on building a CAD system. Such a system is then used to distinguish benign input images from malignant images. A common step in CML and DL is the preprocessing of histological images. For example, the histological images can be de-noised or color normalized [49, 123]. De-noising and color normalization can be helpful steps to eradicate or decrease the deviation of color or contrast information in a histopathological image. The deviations may be caused by staining of a tissue, or by nonstandard scanning condition of a histological slide. Subsequently, the quality of histopathological images can be improved, which assists in capturing coherent characteristics of the content of the tissues. Coherent characteristics increase the effectiveness of the classification model.

After pre-processing a histopathological image, the feature extraction step follows. Such a step attempts to find a robust set of descriptors. Commonly, the attributes are computed at the tissue level. Texture-based techniques (e.g., LBP, GLCM, and others) are proposed in the literature [103, 28] to distinguish normal from abnormal tissues. However, the feature extraction from a histopathological image can be done at the cellular level, and in this case, applying color deconvolution can be of great use in obtaining H & E components to highlight cellular components [222]. Some studies propose first to segment the histopathological image, and then extract attributes from those segmented regions of interest [64, 101]. Other methods that are inherently different from texture-based approaches are proposed, e.g., visual perception-based features [103] which try to mimic how the human visual system differentiates the patterns in histopathological images. Subsequently, after hand-engineering robust and discriminative features, the choice of a CML technique (e.g., SVM or MLP) should have a minimal effect on the accuracy and efficiency in classifying a histopathological image.

The other adopted pipeline is DL which is considered the state-of-the-art approach for machine vision tasks [104]. A particular model of DL is frequently used for histopathological image classification, namely CNN, e.g., in [101, 181]. As such, the CNN model takes the raw histopathological images as input and applies a series of operations (e.g., convolution, pooling, batch normalization, etc). Therefore, the use of such operations in designing a deep neural network model plays a very important role in the robustness of classifying the images. Thereafter, the learned feature maps are fed to a classification layer, typically

SoftMax. Instead, some studies make use of the DL model to be utilized as a feature extractor. Then, those feature vectors are fed to other CML models [32, 76].

However, DL models tend to achieve worse results when trained with a limited number of samples, especially when trained from scratch [226]. There is a limited number of samples of well-annotated medical images (as shown in Table 2.1). In this case, DL models might potentially suffer from generalization and over-fitting. Once the model overfits the datasets, it will not be able to generalize when it is presented with new samples. To overcome this problem, many studies propose to apply data augmentation and fine-tuning models that are pre-trained on a different task [226, 217].

2.4.1 Tissue Classification

The histopathological images are typically used to identify and characterize the stage of a complex tumor. A histopathologist examines a histologic slide visually seeking particular patterns to realize the diagnosis for a patient [165]. Similarly, a considerable number of computational techniques are devised to mimic the process of a histopathologist for analyzing histopathological datasets to build an automated framework to classify normal vs. abnormal tissues. Given the properties of histological images, some studies converting the images to different color spaces [134], or applying color deconvolution [222] to obtain H & E components that highlight the cellular components. Others, applying color normalization [123, 33] which helps to avoid drastic color variations. In fact, another reason might be for a stain normalization that some samples of a dataset are collected from different locations, where the protocol of staining might differ; therefore, it is essential to apply stain normalization prior to training a model for enhancing a classification performance. Other problems specific to histopathological image classification are discussed by Komura et al. in [108].

In Section 2.2.1 publicly available histopathological datasets have been summarized. In the following subsections, we present a summary of CAD systems where we divide our summarization as: 1) Traditional Learning-Based Methods, and 2) Deep Learning-Based Methods. Particularly, we report the summarization of competing techniques that evaluated their algorithms based on the publicly available datasets. Further, we also report their corresponding performance. Later, we concluded the ‘Tissue Classification’ section by providing a discussion.

Traditional Learning-Based Methods

First, we provide a summary of methodologies with their corresponding classification accuracy that proposed traditional descriptors. A considerable number of these traditional descriptors are based on texture features. As such, those texture-based descriptors can be computed either directly on the gray-level or applied on the directional image representations obtained using a spectral method. An advantage of using a spectral method is the characterization of objects in a histopathological tissue with coarse- and fine-grain directional sub-bands. Therefore, a histopathological image can be analyzed using various

resolutions and orientations in which structural information (i.e., step edges) are well localized. This has the potential to lead to an effective classification model to differentiate the various tissue types. For instance, Ribeiro et al. [177] proposed using a fusion of descriptors that were computed from both spatial and curvelet coefficients (i.e., fractal measures and Haralick features). A subset of the features was selected using RELIEFF and used to train a Polynomial classifier (PL). Ribeiro et al. examined their proposed method on two datasets. When it was evaluated on Kather’s dataset to classify stroma vs. tumor using only 250 examples of each class, they reported an accuracy of 97.68%. They reported an AUC value of 0.994 for classifying Warwick-QU. Similarly, Bruno et al. [28] applied LBP but only on the curvelet coefficients that were extracted from the UCSB dataset. Then, ANOVA feature selection was used to reduce the dimensionality. Finally, a PL was used as a classifier which led to an accuracy of 100%.

In this thesis, we proposed to extract textural features (i.e., GLCM, LBP, LOSIB, and SFTA) but in the shearlet domain where both magnitude and phase coefficients were utilized. Our feature representation was reduced using PCA, and was examined on four public datasets. The best-reported accuracy when an SVM model was trained on multi-class Kather was 92.56%, on BreakHis when all magnification factors were treated as one dataset the accuracy was 91.73%, on Epistroma was 98.04%, and on Warwick-QU was 96.29%.

In addition, Kather et al. [103] proposed to combine various hand-crafted features to potentially boost the classification results: lower-order and higher-order histogram features, LBP, GLCM, Gabor filters, Perception-like (i.e., coarseness, contrast, directionality, line-likeness, and roughness) features. When a subset of descriptors was used with an SVM model to classify Kather’s dataset, the highest reported accuracy was 98.6% to classify stroma vs. tumor. However, in the context of the 8-class classification problem, the highest classification accuracy reported was 87.4%.

Other studies also proposed texture-based descriptors while incorporating the color information. For example, Bianconi et al. [22] proposed a modified version of LBP which considered intra- and inter-channel descriptors in combination with local color contrast. The authors tested their descriptors with a nearest-neighbor while using different datasets. The highest reported accuracy while using multi-class Kather’s was 93.4%, while using Lymphoma was 89.5%, while using Epistroma was 97.3%, and while using BreakHis the image-level classification accuracy range was from 91.9% to 93.8% for the various magnification factors. Similarly, Di Ruberto et al. [47] proposed to utilize each color channel independently. As such, Di Ruberto et al. proposed extracting three sets of features: GLCM features from the red channel, gray-level run-length matrix (GLRLM) features from the green channel, grey-level matrix (GLDM) features from the blue channel. A feature selection technique was applied for each set of features, and then combined and fed to train an SVM model. Although Di Ruberto et al. consider only intra-channel features (i.e., each channel R, G, and B was utilized separately), their model was able to achieve high accuracy of 96.0% on the HistologyDS dataset, and accuracy of 96.8% on the Lymphoma dataset.

Orlov et al. [161] applied a color deconvolution on the Lymphoma dataset to obtain

the H & E components. The authors proposed using transform-based approaches (i.e., wavelets, Chebyshev, and Fourier) for each component. Then, features were computed and combined to form a single raw vector of histological images of the Lymphoma dataset. However, the best classification results were produced while utilizing the eosin components only. It is an indication that transform-based techniques are capable of capturing features at different locations and at multiple resolutions. When Orlov et al. used RBF, they reported an accuracy of 99%. Further, Gonzalez et al. [71] studied the Lymphoma dataset and reported perfect accuracy. Gonzalez et al. proposed utilizing orthogonal moments (i.e., DSMs) that are scalar and uncorrelated. A sliding window was applied to extract those moments locally. Also, N-way ANOVA was used to check the statistical significance of the descriptors and window size.

Other generic traditional descriptors were also considered in the literature for the classification of histopathological tissues. For example, Spanhol et al. [204] and Yu et al. [235] both examined several hand-crafted features for breast cancer tissue classification. Those examined features, for example, GLCM, LBP, CLBP, LPQ, Oriented FAST and Rotated BRIEF (ORB), Threshold Adjacency Statistics (PFTAS), morphological features, and SURF. These generic traditional descriptors showed promising classification performance while classifying the breast cancer tissues. It is also worth noting that combining contrasting descriptors, e.g. in [148], can boost the performance of a classification model. While traditional ML-based techniques showed promising classification performance, they require several steps starting from pre-processing, feature extraction/selection, and classification. However, traditional ML-based techniques are capable of handling the scarcity of histopathological tissue samples. With a low number of samples of given tissue type, they are also less susceptible to over-fitting in comparison with DL-based techniques.

Deep Learning-Based Methods

On the other hand, better performance can be achieved using DL, and at times the combination of hand-crafted and deep features boosts the accuracy of a classifier. For example, Alom et al. [11] proposed using the Inception Recurrent Residual Convolutional Neural Network (IRRCNN) model with data augmentation which was trained on BreakHis data for each magnification factor. Alom et al. reported an image-level and patient-level accuracies of almost 98% for each magnification factor. Similarly, Xie et al. [230] reported an image-level and patient-level accuracy of almost 100% for each magnification factor of the BreakHis dataset while fine-tuning Inception ResNet-V2 with data augmentation. Furthermore, Jiang et al. [99] proposed to use SE-ResNet, a model based on the combination of the Squeeze-and-Excitation block. When the model was used to classify the BreakHis dataset, the reported accuracy at the image-level was in the range of 98.87% to 99.34% for the various magnification factors. Moreover, Xu et al. [232] first used a decision DL model. As such, the model aimed to decide whether to crop the original image of BreakHis images and whether the cropped patch was necessary for the classification stage. Then, the selected patches were used to train another model for the classification, which in turn provided feedback to the decision network. Such an approach led to an image-level classification accuracy of approx. 98% for each magnification factor.

Wang et al. [222] reported an accuracy of 92.6% while utilizing only 125 images of each class from the multiclass Kather’s datasets to train a bilinear CNN model. However, higher accuracy of 96.88% for classifying multi-class Kather’s dataset was reported by Maguolo et al. [136]. Maguolo et al. re-trained VGG16 models with different activation functions (i.e., Rectified Linear Units (ReLU), Mexican ReLU, S-Shaped ReLU, and Piecewise Linear Unit). Then, the outputs of these models have ensembled using the sum rule. However, in this thesis, we proposed to utilize deep learning models as unsupervised feature extractors. We reported an accuracy of 95.40% while using only a single DL model, i.e., descriptors that were extracted from DenseNet-201, to classify the multi-class Kather’s dataset.

In the case of classifying the BACH dataset, a 4-class classification problem, many successful techniques utilized DL models. A common approach was to train a DL model based on patches to increase the number of samples [181]. Another approach was to fine-tune pre-trained models that were successful in machine vision tasks [217]. An appropriate technique to conquer the difficulty of building an accurate classifier was to apply a hierarchical classification model. First, the images were grouped into carcinoma and non-carcinoma to train a classification model. Thereafter, a binary classifier was used to find the specific label (i.e., in situ or invasive) in the case of carcinoma. Similarly, the same approach was applied when an image was classified as non-carcinoma. An example of the aforementioned approach is by Kone et al. [109] who fine-tuned ResNeXt50 with data augmentation. Kone et al. reported an accuracy of 99% on the BACH 2018 dataset. However, such a hierarchical approach is computationally demanding. Instead of dividing and conquering, Iesmantas et al. [94] proposed to train a custom shallow CNN model of 5-layers with two more capsule layers. This model was trained based on extracted patches from the BACH dataset, and the final label of an image was predicted using majority voting which led to an accuracy of 87%.

Kumar et al. [114] examined traditional (i.e., LBP and bag-of-visual words) and deep features techniques on the Path960 dataset which contains 20 labels, but the highest classification results were achieved using Bag-of-visual words. Similarly, Alhindi et al. [4] investigated traditional (i.e., LBP and the histogram of gradients) and deep features on Path920, but LBP descriptors led to their highest classification results. Kumar et al. reported comparable results while using deep features with 94.72% accuracy using VGG16. Alhindi et al. reported a low accuracy (i.e., 81.14%) using VGG19. However, it is worth noting a study by Alom et al. [10] who proposed to use IRRCNN with different data augmentation techniques (i.e., rotation, shear, zooming, and horizontal and vertical flipping). Alom et al. [10] reported an accuracy of 98.79% while classifying Path960.

Alternatively, DL models (e.g., CNN) can be used as a feature extractor. However, such models produce high dimensional descriptors from a particular layer. Combining feature maps from multiple layers extracted from a DL model yields even a higher dimensional space of features. Therefore, many authors proposed using a dimensionality reduction [32, 76, 200] to speed-up the training process of a classifier. For example, Song et al. [200] proposed not only to use deep features, but also utilized hand-crafted features. Song et al. examined the following features: Dense SIFT, multiple layers of RBM, features of fine-tuned CNN (i.e., VGG-VD) model. Then, those features were encoded using the FV. Finally, a separation-guided technique was used to reduce the dimensionality. Particularly,

CNN-based encoded descriptors using FV gave the best accuracy. Song et al. [200] reported an accuracy of 98.3% using an SVM model to classify UCSB.

Discussion

Remarkable performance has been achieved in the literature (also see Appendix A for summarization of other studies) utilizing those publicly available datasets, yet further advancement can be made in the histology image classification. Several variables are often not taken into consideration, for example, 1) the split of the training/testing is different across different studies. Also, 2) some authors are using a subset of the data which might influence the generalization of the model being trained. Although logically using fewer samples to training a model is appealing, it is crucial to use representative and challenging samples to train and test a classifier to obtain a robust model. 3) Some studies also rescale the histopathological images; however, if training a deep learning model with variable size images, then it is a must to rescale all images as the model expects a fixed size of images for retraining (or training from scratch).

Therefore, in Chapter 3, we present novel shearlet-based hand-engineered texture descriptors to classify tissue types (particularly of four datasets: Kather, BreakHis, Epistroma, and Warwick). As such, statistics are computed from each directional sub-band of the complex shearlet domain, where not only the magnitude but also the phase coefficients are utilized in our study. Unlike other studies that use the shearlet transform, e.g., He et al. [81], Dong et al. [50], and Zhou et al. [241] where the phase coefficients are discarded. The phase coefficients are as significant as magnitudes for their embedded directionality information. Textural descriptors can capture local (or global) statistical properties [42, 198]. However, applying such methods directly in the spatial domain images can potentially lead to missing essential characteristics of histopathological tissues due to the complexity of such images. Instead, we leverage both the magnitude and phase coefficients to compute various statistics from the directional sub-bands which can potentially lead to capturing significant information. Furthermore, in our investigation, we set our experiments using common testing techniques in the literature for a fair comparison. As evident, our model can generalize on four publicly available datasets while accomplishing classification performance on par with state-of-the-art techniques.

In the case of deep learning strategies, with the scarcity of well-curated histopathological datasets for training/testing deep neural networks [171, 216], training a deep learning model can be a challenging approach. Some of the best-achieving methods are based on training a DL model from scratch or opting to investigate a shallow DL model [101, 205]. Such techniques might not be feasible because aggressive data augmentation and a long training process are required. Therefore, we instead have explored the use of DL models in two contexts for histopathological image classification (as described in Chapter 4): 1) pre-trained models are examined as unsupervised feature extractors, 2) pre-trained models are used as weight initializers either by fine-tuning weights of all layers or freezing some layers. Subsequently, our strategies are requiring drastically less time for training and overcoming the susceptibility to over-fitting [240]. Our deep strategies lead to high classification results on four publicly available datasets (i.e., multi-class Kather, BreakHis, Epistroma,

Warwick-QU). Our expectation that our techniques can be extended to biomedical image classification that exhibits similar complex patterns. Therefore, we would like to examine our hand-crafted and deep learning approaches on a different modality that present an unrelated challenge, namely MRI. MRI samples are large because of the number of slices per brain scan.

2.5 Early Diagnosis of Alzheimer’s Disease from MRI

In-depth reviews of existing techniques for AD are provided by Tanveer et al. [214] and Ebrahimighahnavieh et al. [56]. According to Ebrahimighahnavieh et al. MRI data for AD classification can be handled in one of the following ways:

- **Voxel-based methods:** this approach looks at the voxel intensity values from the whole brain MRI or in the tissue components (i.e., in the gray matter or white matter of the MRI by applying voxel-based morphometry).
- **Sliced-based methods:** the problem of the 3D volume is reduced to 2D where each slice can be treated as an independent image without looking at the adjacent slices.
- **Region of Interest (ROI)-based methods** this approach requires prior knowledge of the normal regions and identifying the damaged regions in the MRI scans. Therefore, the focus of this approach can be on the damaged regions.
- **Patch-based methods:** this approach partitions an image into patches. This approach can be sensitive to the patch size, but does not require prior knowledge about the regions in the brain like in the ROI-based methods.

However, in this section, we divide our review of techniques that used the ADNI and OASIS data repositories (see Section 2.3 for more details about the repositories) for Alzheimer’s disease classification into two categories: traditional and deep-learning-based techniques. Such techniques are relevant to our proposed method of AD classification (see Chapter 5). Finally, we discuss how our previous techniques are extended to be used on a fundamentally different type of dataset, namely MRI.

Traditional Learning-Based Methods

Previously, we reviewed and discussed CAD systems that are related to histopathological image classification. The same principle steps apply to the MRI data, but the MRI scans can be handled differently (e.g., utilizing a sliced-based method).

An example of a study using a sliced-based technique is conducted by Altaf et al. [12] who proposed to combine both texture-based and clinical data for AD classification. The bag of visual word technique was utilized with a gray level co-occurrence matrix, scale-invariant feature transform, local binary pattern, and histogram of gradient. The texture-based features were extracted either from the whole image or segmented regions, i.e., grey

matter, white matter, and cerebrospinal fluid. However, when GLCM was obtained, four texture descriptors were computed, i.e., contrast, correlation, homogeneity, and entropy. It is worth noting that only a subset of slices was used. If all slices were used, then the feature vectors were averaged.

Other studies compute features from slices and/or voxels. For example, Nanni et al. [147] proposed the fusion of texture-based and voxel-based features to train three binary SVM classification models independently: AD vs. CN, pMCI (mild cognitive impairment patient that might convert to AD) vs. CN, and pMCI vs. sMCI (mild cognitive impairment patient that might not convert to AD). Alternatively, Faturrahman et al. [59] proposed using gray matter maps to extract 1) mean and standard deviation of voxel values, and 2) segmentation values of the voxel location. Such attributes were used to train a deep belief network.

In addition, Salvatore et al. [184] proposed training an SVM based on feature extraction and reduction of MRI samples using descriptors from PCA. PCA descriptors then were sorted in a descending order based on their Fisher Discriminant Ratio (FDR) scores. Salvatore et al. applied three classification tasks with the following: AD vs. CN, pMCI vs. CN, and pMCI vs. sMCI.

Other studies proposed to extract information from ROI. For example, Lee et al. [122] proposed using GLCM but instead of applying this texture-based technique on the spatial domain slices, the authors computed the co-occurrence matrices from the voxel pairing. Such attributes were computed from different regions of MRI scans, i.e., the hippocampus, precuneus, and posterior cingulate cortex.

Some studies propose to use spectral methods with the MRI data for multi-scale and different orientations analysis. For instance, Jha et al. [98] proposed using a dual-tree complex wavelet transform (DTCWT) to transform each slice of an MRI. These wavelet coefficients were reduced using PCA. Finally, a feed-forward neural network was used to classify AD from CN samples. Similarly, Feng et al. [61] proposed using Daubechies wavelet transformation for each scan of an MRI sample. From the directional sub-bands, energy regions of interest were obtained to compute the energy descriptors, which were then fed to the nearest neighbor classifier for AD classification.

Deep Learning-Based Methods

Unlike the traditional ML techniques, deep learning is utilized to handle the complexity of MRI data in two ways using 2D or 3D deep learning models. An example of a sliced-based DL model was proposed by Jain et al. [97]. Jain et al. fine-tuned a pre-trained VGG-16 where each MRI slice was considered as an input by shaping each slice as $h \times w \times 3$ to fit the input layer of the CNN model. Each MRI sample was pre-processed where a subset of slices was selected based on image entropy. Similarly, Yagis et al. [233] explored fine-tuning VGG-16 and ResNet-50 for AD classification while each slice of an MRI was considered as an independent labeled example. They only fine-tuned the last three layers and other layers were frozen. Also, some slices were discarded based on image entropy.

The split of slices is often done randomly where slices from the same patient can appear in the training and the testing dataset. Arguably, subject-based splitting where slices of a patient can only appear in the training or testing split is more appropriate. Yagis et al. concluded that the subject-based splitting, i.e., slices of an MRI sample either appear in the testing or training set exclusively, makes it more difficult to obtain high accuracy.

Instead of fine-tuning a DL model, Wang et al. [223] proposed to train an 8-layer CNN with data augmentation for the classification task of AD vs. CN. Prior to training the model, MRI scans were pre-processed for spatial normalization, smoothing, slice selection, and histogram stretching. Similarly, Abrol et al. [1] proposed training a ResNet model first to predict the MCI class only. Afterwards, this model was fine-tuned to predict AD vs. CN samples. However, Lian et al. [125] proposed a patch-based hierarchical convolutional network, where various data augmentation techniques were applied, i.e., random flipping, distorting, and shifting, for the following binary classification tasks: (AD vs. CN) and (progressive MCI vs. stable MCI).

Similar to Lian et al., Mendoza et al. [141] proposed a patch-based autoencoder classification model to distinguish AD vs. CN. This autoencoder was trained on a single 2D slice and three independent views were explored: Axial, Coronal, and Sagittal anatomical planes. The final label of a 2D slice was obtained using a majority voting rule from the ensemble of patch classifications. As well, Pan et al. [162] proposed training a SENet (i.e., Squeeze-and-Excitation Network) with data augmentation. As such, a various number of CNN models were trained on different views, i.e., sagittal, coronal, or transverse MRI slices. Then, based on the particular binary classification task, the outputs of these various CNNs were ensembled based on the three-axis slices.

Cao et al. [31] proposed first computing the gray matter images from the MRI data. From such images, only a subset of the MRI slices was selected based on image entropy. Further, the authors propose a framework that had two pipelines in which the first pipeline processed the images as 2D, and the other pipeline processed the images as 3D. Then, the feature maps from both pipelines were fed into a SoftMax layer to apply the following classification task independently: (MCI vs. CN) or (MCI vs. AD).

Discussion

Conventional and deep learning techniques are proposed for AD classification from MRI data. Different key factors play an important role while building such an automated framework to classify AD; hence, it is difficult to compare one study to another. These factors are the MRI data and selected participants, the image pre-processing techniques, the split of the dataset to train a model, and the classification task. Consequently, to assess and compare the performance of a method objectively is difficult because it is not clear which part of the technique led to a tangible enhancement.

A popular first step before applying a conventional or deep learning approach is to pre-process the MRI samples (e.e., skull stripping, intensity normalization, or other pre-processing techniques). Software packages like FreeSurfer or CAT12 are popular for pre-processing MRI data [58]). However, most of the previously summarized papers do not

explicitly list the exact patient IDs of the ADNI or OASIS dataset that are used in their studies. Further, the split of the training and testing data is different. One major key component that also influences the generalization of a CML or DL model is mixing the MRI slices for training and testing of a classification model. For instance, while labeling each slice and randomly splitting these slices for training and testing, slices of a patient might be in both sets of training and testing; hence, a high correlation of some examples exist between both sets. Therefore, the reported classification results might be very optimistic [233, 141]. If this happens, then there was an ‘information leak’. Therefore, one must make sure to strictly split the data where a subject appears exclusively in the training or the testing split only.

Therefore, for fair comparison and for which components that are improved of our proposed techniques, we set our experiments as close as possible to previous work published by Salvatore et al. [185], and Hon and Khan [85]. In this manner, we use the same MRI data and use a similar train/test protocol to examine our novel proposed techniques to classify AD (as described in Chapter 5). Unlike many studies that proposed using texture-based techniques [122, 12], our hand-crafted technique is adjusted to use the 3D shearlet transform instead of the 2D shearlet. The 3D shearlet transform considers all slices while transforming the volumetric MRI sample. To overcome the problem of the high-dimensionality of a feature vector per MRI sample, each slice is represented by the average presence of shearlet coefficients. Then, a texture-based method is used to summarize those coefficients for each slice. Further, we also propose to fine-tune a DL model while feeding stacked slices of an MRI. Then, we leverage such a fine-tuned model as a feature extractor to train a CML model. Unlike other studies, for example [97, 233], the authors in these studies fine-tune a CNN model based on each slice independently. Also, some studies [125, 162] trained a CNN model from scratch with the need for data augmentation. Thereafter, given that shearlet-based and deep features calculate attributes from the MRI differently, we concatenate both shearlet-based descriptors with deep features. We apply two strategies for training a classification model after concatenating the feature vectors. In one strategy, we feed the concatenated descriptors to an SVM or DTB, in the other, we use a custom CNN model which will further process the attributes and then obtain the classification model.

2.6 Summary

In this chapter, we have summarized CAD techniques that can be categorized as traditional and deep learning-based techniques for medical image classification. Classical texture-based approaches are used extensively to distinguish between different pathological conditions in medical images. In addition, it has been shown that DL techniques and sometimes a combination of hand-crafted descriptors with deep features can lead to higher accuracy in classifying different medical images.

Although DL techniques have received increasing attention in recent years, yet the application of DL in medical image diagnosis is under scrutiny [57]. One possible reason is the lack of standardized protocols, for example, data acquisition, or training and testing

classification models. Another significant reason is that DL models require an abundance of well-annotated samples in order to be able to learn the hidden patterns in medical images. These datasets need to be well-annotated for supervised learning models, which is a time-intensive task. The difficulty to obtain a large representative population of medical images limits the application of DL and its advancement; hence, it is challenging to build a model that is being able to generalize to new samples. Given the fact that there is a shortage of well-annotated datasets, a desirable approach to overcome such obstacles is to use a shallow DL model, to fine-tune pre-trained models [33, 94, 70] instead of training a model from scratch, adopting semi-supervised [242, 16], or unsupervised [238, 2] techniques for medical image classification.

Nevertheless, we model the classification of histopathological images by proposing, 1) texture-based features in the shearlet domain that capture directionality not only in magnitude coefficients but also in phase coefficients. Furthermore, 2) we utilize pre-trained CNN weights for unsupervised feature extraction. Also, we investigate weights freezing of shallower modules of SqueezeNet and ResNet, which potentially reduces the time required for re-training the model. Further, we propose exploring whether we can extend the descriptors that we will develop for histopathological image classification to classify Alzheimer’s disease from MRI samples.

Chapter 3

Shearlet-based Techniques for Histological Image Classification

3.1 Overview

A number of imaging modalities are available, e.g., magnetic resonance, x-ray, ultrasound, and biopsy, where each modality can reveal different structural aspects of tissues. However, the analysis of histological slide images that are captured using biopsy is considered the gold standard to determine whether cancer exists [95]. Furthermore, it can reveal the stage of cancer. Supervised machine learning can be used to classify histopathological tissues. A number of computational techniques have been proposed to study histopathological images with varying levels of success. Often handcrafted techniques based on texture analysis are proposed to classify histopathological tissues.

In this chapter, we first provide a brief related work (see Section 3.2) of studies using spectral techniques to classify the histopathological dataset. We construct a novel feature space to automate the classification of tissues in histology images as described in Section 3.4. Our feature representation is to integrate various feature sets into a new texture feature representation. All of our descriptors are computed in the complex shearlet domain. With complex coefficients, we investigate not only the use of magnitude coefficients, but also study the effectiveness of incorporating the relative phase (RP) coefficients to create the input feature vector. In our study, four texture-based descriptors are extracted from the shearlet coefficients: co-occurrence texture features, Local Binary Patterns, Local Oriented Statistic Information Booster, and segmentation-based Fractal Texture Analysis. Each of these descriptors captures significant local and global statistics. Therefore, we study them individually but additionally integrate them to boost the accuracy of classifying the histopathology tissues while being fed to classical classifiers. To tackle the problem of high-dimensionality, our proposed feature space is reduced using principal component analysis. In our study, we use two classifiers: support vector machine (SVM) and Decision Tree Bagger (DTB) (described in Section 3.3).

3.2 Related Work

The early identification of cancer is crucial for the histopathologist to propose an appropriate treatment for patients. The process of histopathological tissue classification is tackled in different ways. In this section, we focus on studies that use spectral techniques. For instance, Kather et al. [103] and Linder et al. [126] proposed the use of multichannel Gabor filtering to classify histology images.

Bruno et al. [28] proposed applying LBP on the curvelet coefficients of the transformed image. In comparison to Bruno et al., we also use descriptors computed from a directional wavelet transform, but we demonstrate that it is advantageous to integrate various descriptors computed in the shearlet domain to create the feature space.

A closely related technique to ours is proposed by Ribeiro et al. [177]. Ribeiro et al. computed descriptors with fractal techniques and Haralick descriptors from both spatial images and curvelet coefficients to classify colorectal histology tissues. In contrast to Ribeiro et al., we utilized both magnitude and phase coefficients of the complex shearlet domain. A similar approach to ours is also proposed by Vo et al. [220] who extracted both, phase and magnitude descriptors for textured image retrieval, but applied to non-medical texture image samples.

Other studies also proposed using the shearlet transform which is a truly anisotropic form of a wavelet. It has been previously used in different studies in which only the magnitude coefficients are utilized for the classification of textured images. He et al. [81] classified textured images by using shearlet-based descriptors. They construct local energy features from the shearlet coefficients, quantize and encode these local energy features making the representation rotationally invariant. All energy histograms of all decomposition levels are accumulated to represent the image characteristics. Instead, Zhou et al. [241] utilized only specific levels of the decomposition of the shearlet domain for breast tumor ultrasound image classification.

Dong et al. [50] proposed an approach for textured image classification and retrieval, where the dependencies of adjacent shearlet subbands are modeled using linear regression. To represent the shearlet subbands for classification, energy descriptors are computed. However, the textured image retrieval consists of both statistics in the contourlet and shearlet domains.

Meshkini and Ghassemian [142] used the inner product of co-occurrence matrix and magnitude coefficients of shearlet transform to classify textured images. Different from many published studies, we use both magnitude and phase coefficients of the shearlet transform in our work. Our proposed system is based on the combination of the discrete complex shearlet transform with co-occurrence, LBP, LOSIB, and SFTA which to the best of our knowledge has not been investigated before. In particular, we consider different scales of magnitude and phase coefficients in the shearlet transform, as we describe in Section 3.4.

3.3 Classification Algorithms

There exist many classifiers. We have examined a number of them in our study, e.g., naive Bayes classifier [66], MLP [80], discriminant analysis [63], nearest neighbor [43], and others. As well, there are other classifiers that are based on weak learner ensemble, e.g., adaptive boosting (AdaBoost) [65] and gradient boosting [35]. In this study, we also examined AdaBoost, but gradient boosting would be worth to examine in the future. AdaBoost and gradient boosting are based on ensembling weak learners in a sequential manner. The new weak learner, i.e., decision tree, in AdaBoost puts more weight on previously misclassified examples. As such, the algorithm of AdaBoost attempts to reach a model that can correctly classify these examples. Subsequent alternatives to AdaBoost have emerged, e.g., gradient boosting, which is also based on an ensemble of weak learners. In comparison to AdaBoost, gradient boosting identifies the limitation of the previous weak learner by the gradient. As such, gradient descent is utilized to minimize the loss of the current weak learner in comparison to the previous weak learner.

In this study, we particularly analyze the performance of two classifiers, which we found to work well with our classification tasks. In particular, we consider one of widely used classifiers in biomedical research, i.e., an SVM classifier. In addition to SVM, we consider one type of decision tree, which is a DTB [26], that has the trait of being non-parametric. As such, the distance between feature vectors is not computed in constructing decision trees. In contrast, SVM performance is based on kernels that can influence the classification performance [236]. It is also worth noting that different classifiers were compared by Thanh and Kappas [215] who demonstrated that the highest classification results were obtained by an SVM model, and then were followed by decision trees. In this study, we also use SVM and DTB. Both are briefly described below.

3.3.1 Support Vector Machine (SVM)

In our study, we utilize an SVM with pairwise classification (i.e., one-versus-one class decisions) [79]. Prior to training, we normalize all feature vectors to have equal mean and variance. The kernel function transforms the input data into a higher-dimensional feature representation, from which a hyperplane is formulated for classifying the input tissue dataset.

SVM training requires the solution of a quadratic programming (QP) problem. To simplify the solution of the problem, a sequential minimal optimization (SMO) solver is used in our study [169]. This solver simplifies the QP into a smaller series of QP for the training of SVM.

SVM training is impacted by the choice of the kernel function. In our study, we choose a radial basis function as a kernel function as it universally approximates the training dataset accurately. Therefore, the regularization parameter (C) is the only remaining parameter that can be optimized. In our attempt to choose the best value of C , trials of values between [1, 5] with an increment of 1 were conducted. We found that in general $C = 5$ delivered the best classification performance. Therefore, the value of $C = 5$ is set for all of

our experiments. Most importantly, C is a regularization parameter that manages the penalty on the margin-violating support vectors, and helps in avoiding over-fitting. When a smaller penalty value is chosen, the wider the margin, and typically more support vectors lie within the margin. On the other hand, a larger penalty value might lead to fewer support vectors on the margin. A larger penalty value can lead to heavily weight misclassified support vectors, a smaller margin, and a higher risk of over-fitting. More details about SVM can be found in [66].

3.3.2 Decision Tree Bagger (DTB)

DTB [26] is an ensemble classifier, i.e., each classifier in the ensemble is a classification and regression tree (CART). A decision tree splits the dataset. The impurity of the dataset is reduced by each split in comparison to the dataset being split. The Gini index [132] is utilized to measure the impurity of each dataset. In the decision tree, the original dataset is associated with the root node, and each child node holds a partition of the dataset of the root. Partitioning the dataset is repeated at each of child nodes. The goal of such a process is to reach a pure partition of the data at the leaf nodes for each branch of the tree. When a tree is built from the training dataset, a tree can be traversed from the root to each of the leaf nodes to generate rules. Afterward, these rules can be used to classify the unknown dataset. It is worth noting that the tree structure can be different for the same classification problem depending on the training dataset. As a result, the classification results may also be different.

DTB grows a number of decision trees in an ensemble. In our experiments, we set the number of grown trees to be 750 decision trees. The outcome is combined using bootstrap-aggregated (bagged) decision trees. As such, replicas of the training set are generated to train individual decision trees with replacement from the provided training set. These replicas vary, which generates different trees, and the outcome of these trees can be bagged to obtain an improved classification result. As an advantage, the bagging technique reduces the susceptibility to over-fitting. In our utilization of DTB, the grown decision trees select a random subset of attributes at each node to determine the split as in a Random Forest [27]. The size of the random subsets is the square root of the number of attributes in our experiments. As an advantage, this can potentially speed up the process of building a decision tree.

3.4 Method

Figure 3.1 provides an overview of our proposed system for the classification of histopathological images. A detailed description of each component of our method is provided in the following sections. MATLAB® 2017b is utilized for the implementation of our techniques.

Our proposed method consists of three steps:

- Step 1: For a given histopathological image, we apply the complex shearlet transform. With the complex coefficients, we compute the magnitude and relative phase (RP).

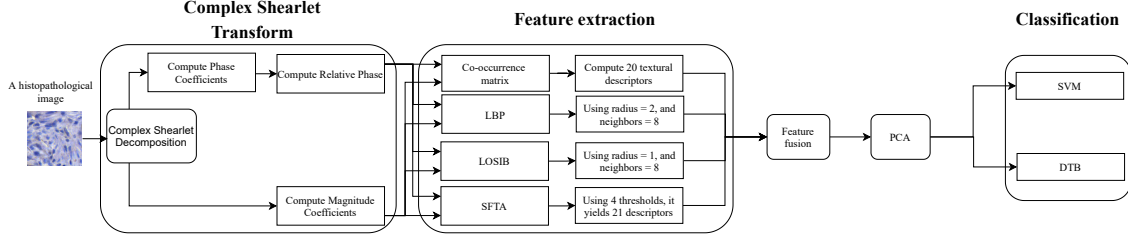


Figure 3.1: Overview of our shearlet-based system for histopathological image classification. © Copyright IEEE 2019.

- Step 2: We then extract four sets of features from the directional sub-bands of the magnitude and RP: co-occurrence based texture features, local binary pattern, local oriented statistics information, and segmentation-based fractal texture analysis.
- Step 3: We then apply one of two different classifiers: DTB or SVM.

3.4.1 Attributes from the Complex Shearlet Transform

In this chapter, we construct our proposed novel descriptors in the shearlet domain [115]. The shearlet transform, as shown by the authors, is an efficient tool to capture 2D directional features from an image. The shearlet coefficients localize spatially distributed discontinuities and are contrast invariant [176]. Although curvelets [30] and contourlets [48] have similar properties to shearlets and have been used for image classification [28, 177], both have certain limitations [115]. Curvelet generation based on rotation operator does not preserve the digital lattice, and contourlet arguably lacks a well-established theory.

On the other hand, a characteristic of the shearlet transform is its use of shearing to control directional selectivity. This operation does not destroy the digital lattice, i.e., unlike the curvelet transform which does in most cases. Therefore, the shearing allows a digitized implementation of the shearlet transform. The shearlet transform is based on affine system theory. Such an affine system consists of the following three operators.

1. First, the dilation operator that is based on the parabolic scaling A_a matrix, where $a > 0$:

$$A_a = \begin{pmatrix} a & 0 \\ 0 & a^{1/2} \end{pmatrix}$$

2. Next, to capture the orientations of waveforms, the shearing matrix B_s is utilized, where $s \in \mathbb{R}$:

$$B_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

3. Finally, a translation operator (i.e., $t \in \mathbb{R}^2$) is used as a mean to change the position.

From these three operations, we then can obtain the continuous domain of shearlet where the mapping function is defined as: $SH_{\Psi}f(a, s, t) = \langle f, \Psi_{a,s,t} \rangle$ for $f \in L^2(\mathbb{R}^2)$. Then, the shearlets are computed as

$$\Psi_{a,s,t}(x) = |\det M_{a,s}|^{-\frac{1}{2}} \Psi((M_{a,s})^{-1}x - t). \quad (3.1)$$

Such shearlets capture anisotropic features and generate well-localized waveforms since it uses parabolic scaling and shearing matrices. Both of these operations are embodied in $M_{a,s}$, and defined as follows:

$$M_{a,s} = B_s A_a = \begin{pmatrix} a & a^{\frac{1}{2}}s \\ 0 & a^{\frac{1}{2}} \end{pmatrix}$$

To obtain the shearlets in the discrete domain [55], the continuous shearlets are sampled on a discrete set of parameters. As such, one can choose $a = 2^{-j}$ and $s = -l$, where $j, l \in \mathbb{Z}$, to compute the $M_{2^{-j}, -l}$. Then, $M_{2^{-j}, -l}^{-1} = M_{2^j, l} = B^l A^j$, where B and A can be defined as:

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Given that $j, l \in \mathbb{Z}$ and also the translation parameter t is replaced with $k \in \mathbb{Z}^2$, the discrete shearlet transform then can be obtained as

$$\Psi_{j,l,k}(x) = |\det A|^{\frac{j}{2}} \Psi(B^l A^j x - k). \quad (3.2)$$

Now, having computed the complex shearlet transform, we propose a set of novel descriptors that capture statistical distribution/information of the shearlet magnitude/phase components for each scale and direction [115]. For a complex number $C = x + iy$ where x is the real part and y is the imaginary part, the magnitude and phase coefficients are defined as: $\rho = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}(y/x)$, respectively. In comparison to others [81, 241] who only consider the magnitudes, we investigate the effectiveness of phase coefficients and their ability to classify bio-medical textured images. We empirically show the effectiveness of phase coefficients in Section 3.6.

In a previous study, Vo et al. [220] proposed the extraction of descriptors from the relative phase (RP) of a complex directional filter bank for textured image classification. We adopt this idea, but instead extract the relative shearlet phase coefficients. A given image of size $N \times N$ is decomposed by the complex shearlet transform with S scales and K orientations for each scale. We denote $\theta_{sk}(i, j)$ as the sub-band phase angle coefficient at position (i, j) at scale s and orientation k , where $s = 1, 2, \dots, S$ and $k = 1, 2, \dots, K$. In our study, $S = 4$ and $K = 8$ per scale.

For a given phase coefficient at location (i, j) , RP of a directional sub-band is obtained as follows:

$$RP_{sk}(i, j) = \begin{cases} \theta_{sk}(i, j) - \theta_{sk}(i, j + 1), & \text{if } 1 \leq k \leq \frac{K}{2}. \\ \theta_{sk}(i, j) - \theta_{sk}(i + 1, j), & \text{if } \frac{K}{2} < k \leq K. \end{cases} \quad (3.3)$$

The vertical and horizontal differences are chosen because of the direction of shearing in the shearlet transform. The shearlet transform is multi-scale and multi-directional which provides a multi-resolution analysis of textured images. Statistical descriptors are computed for each directional sub-band (i.e., from magnitude and RP). In our investigation, we use the ShearLab implementation of the complex shearlet transform by Kutyniok et al. [117].

In our research, we explore four different methods to extract informative descriptors from the magnitude and RP of shearlet coefficients. Each method is briefly described below:

Co-occurrence Matrix (CM)

The first feature set is the CM which has been originally proposed by Haralick et al. [77] and then further descriptors were studied [198, 40]. The texture-context information (i.e., from the gray-level images) is represented as a matrix of relative frequencies F_{ij} for two neighboring pixels where the first pixel has a gray level i and the second pixel has a gray level j separated by pixel distance PD . The co-occurrence matrices can be calculated for different orientations, i.e., a matrix contains relative frequencies for quantized angles and distance between neighboring pixels. We utilize a distance of $PD = 1$ and angles $= (0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ)$ to compute the CMs. The extracted descriptors become invariant with respect to rotation by calculating the mean [103] of all four directions for the given distance of PD . We compute twenty textural features: contrast, correlation, energy, autocorrelation, cluster prominence, cluster shade, dissimilarity, entropy, homogeneity, maximum probability, sum of squares, variance, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation, inverse difference normalized, and inverse difference moment normalized. This texture feature extraction from the magnitude/RP coefficients follows common practice; therefore, yielding 20×32 (# of directional subbands) = 640 descriptors for each magnitude and RP. We calculate the CM from the shearlet magnitude and RP components of each directional sub-band. Unlike the spatial domain [77] where the various CMs are concerned with the statistical distribution of the gray tones, the CM of shearlet magnitude and RP components summarize the significant information of directionality features of an image at a particular scale and orientation.

Local Binary Pattern (LBP)

LBP [157] provides rotationally invariant statistics of the local texture occurrence which is typically categorized as ‘uniform’ patterns of the gray tones of an image. However, the magnitude/RP of a transformed image using shearlets captures field dominant direction features [163, 29]. Most importantly, these directionality features can categorize significant structural information (e.g., step edges) [160]. We use LBP to encode the field dominant direction features in a sub-band of the shearlet. Hence, for a given shearlet (SH) coefficient (i.e., representing the magnitude or RP) at the location (i, j) , the LBP of the SH coefficient

is computed as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(SH_p - SH_c)2^p, s(x) = \begin{cases} 1, & x \geq 0. \\ 0, & x < 0. \end{cases} \quad (3.4)$$

SH_c and SH_p are the central magnitude (or RP) values. P are the neighbor coefficients that are surrounding the central magnitude (or RP) values in the circular neighborhood. In this study, a feature vector for each sub-band is computed using a radius of $R = 2$ and a neighborhood of $P = 8$, where the window size is equal to the image size. Based on this formulation, the number of descriptors [157] of each directional sub-band is $(P + 2)$. Thereafter, all sub-band feature vectors of an image are concatenated, obtaining $10 \times 32 = 320$ descriptors for each magnitude and RP.

Local Oriented Statistic Information Booster (LOSIB)

In this study, we propose one more novel technique that combines LOSIB [67] with the shearlet transform to capture intrinsic statistical information from the shearlet magnitude/RP coefficients.

The first step of the LOSIB pipeline is to obtain the absolute difference d_p between the central shearlet value with P neighboring values (i.e., applied to all central magnitude (or RP) values c of a particular directional sub-band) as follows: $d_p(i_c, j_c) = |SH_c - SH_p|$ where $p \in 0, 1, \dots, (P - 1)$.

Thereafter, the differences across the same orientation are averaged according to the following formula:

$$\mu_p = \frac{\sum_{x_c=1}^M \sum_{y_c=1}^N d_p(x_c, y_c)}{M \cdot N} \quad (3.5)$$

where N and M are the height and width of the image, respectively. A feature vector for each sub-band is computed using a radius $R = 1$ and neighborhood $P = 8$. This yields a feature vector of size P for each directional sub-band. Then, feature vectors of all sub-bands are concatenated. This yields $8 \times 32 = 256$ descriptors for each, magnitude, and RP.

Segmentation-based Fractal Texture Analysis (SFTA)

The last set of descriptors proposed in this study is based on SFTA [42]. In the first step, SFTA decomposes the input directional sub-band (of shearlet magnitude or RP) using Two-Threshold Binary Decompositions (TTBD). This results in a number of binary images in which the dimension of the fractal boundaries, the average gray level, and the count of pixels belonging to the region are calculated. In our application of the method by Costa et al. [42], we use a number of thresholds of $n_t = 4$ which yields 21 descriptors for each directional sub-band. Therefore, 21×32 (# of directional subbands) = 672 descriptors for each, magnitude and RP.

Table 3.1: Summary of datasets used in Ch. 3.

DB	#Classes	#Samples	Sample Size
Kather’s	8	5000	150×150
BreakHis	2	7909	700×460
Epistroma	2	1376	Variable
Warwick-QU	2	165	Variable

3.4.2 Fusion of Feature Sets

In addition to comparing the effectiveness of the above four descriptors, we also explore whether an extended feature representation combining different shearlet descriptors could enhance the performance of a classifier. This is justified as different descriptors are conceptually contrasting, and each feature set captures certain aspects. The selected set of descriptors to be used jointly was based on the merit of individual descriptors and their capability of achieving reasonable accuracy. We test the following fusions:

- Fusion #1: (CM + LBP + LOSIB + SFTA) descriptors of shearlet magnitude and/or RP.
- Fusion #2: (CM + LOSIB) descriptors of shearlet magnitude and/or RP.
- Fusion #3: (CM + LBP + SFTA + CM Dot shearlet coefficients [142]) of shearlet magnitude and (CM + LBP + LOSIB) of shearlet RP.

3.4.3 Principal Component Analysis (PCA) for Feature Reduction

In the previous section, we investigate feature fusion. Evidently, with the combination of different descriptors, the number of features increases. Therefore, we apply PCA to the fusion scheme with the highest accuracy. The PCA coefficients are rotated to maximize the orthomax criterion [78], and to obtain a final basis with simple structure [206]. After principal components (PCs) are rotated to maximize varimax, they are used to project the fused features to a decorrelated space.

3.5 Experiments

3.5.1 Datasets

To evaluate our proposed approach in this study, four different histopathological datasets are used, as described in Section 2.2.1. As a reminder, we provide a summary of these four datasets in Table 3.1.

3.5.2 Cross-validation (CV)

In our study, we utilize cross-validation (CV) to obtain robust statistical results, and in order to be able to generalize the results with each classification model. This approach is commonly used in the literature to examine a proposed technique. Therefore, our choice of the number of folds to split a dataset is based on previous studies. As such, the following dataset are divided using 10-fold CV: **multi-class Kather’s** (as Kather et al. [103]), **Epistroma** (as Ramalho et al. [175]), **Warwick-QU** (as Ribeiro et al. [177]).

The datasets are divided into mutually exclusive folds with approximately equal size (i.e., some of the datasets are imbalanced): 9-folds are used to build a model, and 1-fold is used to test the model. The process is repeated 10-times, such that the test set is different each time. Finally, the overall performance metrics are estimated by taking the mean from the tested 10 independently built models. However, the **BreakHis** dataset is split into 7-folds CV for training and testing as in a previous study [101].

Performance Metric

Commonly used performance metrics to examine the classification performance of histopathological image classification are accuracy (ACC) and area under the receiver operating characteristic (ROC) curve (AUC). In addition to ACC and AUC, we also report sensitivity (Sen), and precision (Prec) as was suggested by the reviewers of our journal extension version [7].

The computation of each evaluation metric is calculated as follows:

- $ACC = (TP + TN)/N$. This metric represents the test examples that are correctly classified over all the number of test examples available. As such, true positive (TP) and true negative (TN) are the correct examples, and $N = (TP + TN) + (FP + FN)$, where false positive and false negative are denoted as FP and FN, respectively.
- AUC: This metric relies on the plot of ROC (i.e., false-positive rate ($=FP/(TN+FP)$) on the x-axis against sensitivity on the y-axis). Therefore, the accumulated area under the ROC curve yields the AUC value (i.e., between 0 & 1, where 1 is a good performing model).
- $Sen = TP/(TP + FN)$ measures the ability of a model to correctly predict the positive condition when it is actually positive.
- $Prec = TP/(TP + FP)$ measures the ability of a model to correctly identify positive cases.

3.6 Results and Analysis

We apply the shearlet transform as detailed in Section 3.4 to four different datasets (described in Section 3.5). We compute the RP and magnitude from the complex shearlet

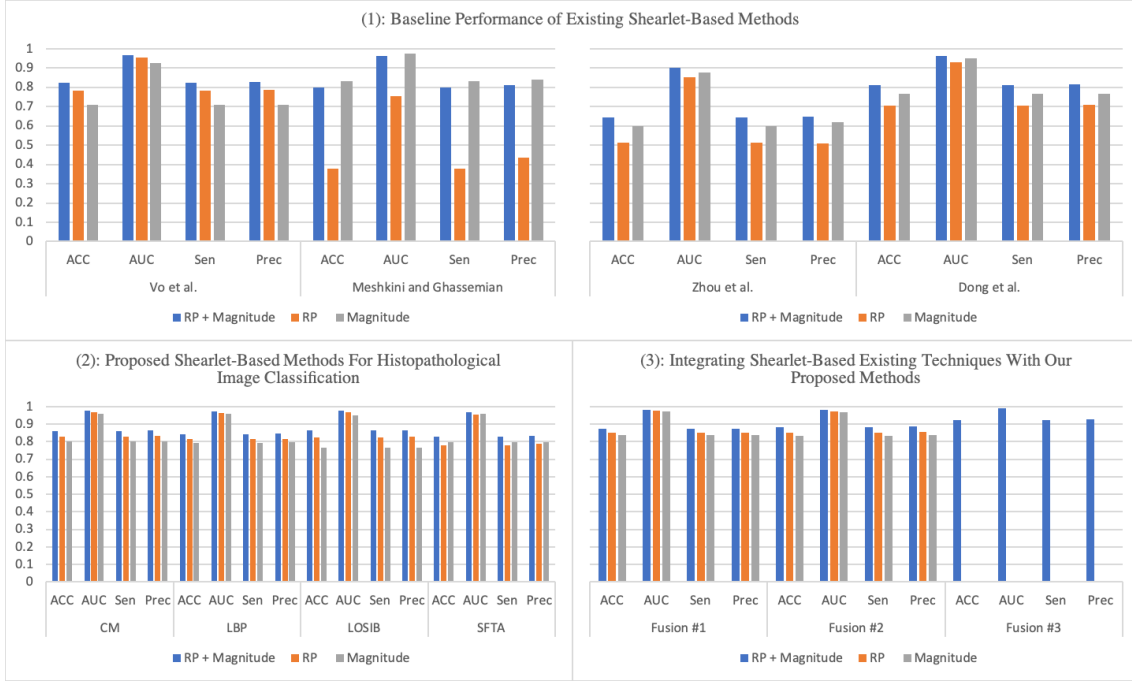
coefficients. Then, to evaluate the strength of our proposed techniques, we utilize four commonly used measures for classification: ACC, AUC, Sen, and Prec.

In the literature, there exist descriptors extracted from multi-directional and multi-resolution wavelets (e.g., shearlet, contourlet, and others). In this regard and to the best of our knowledge, we have implemented descriptors of previously published studies. These descriptors are applied on the transformed (i.e., using complex shearlet transform) histopathological dataset:

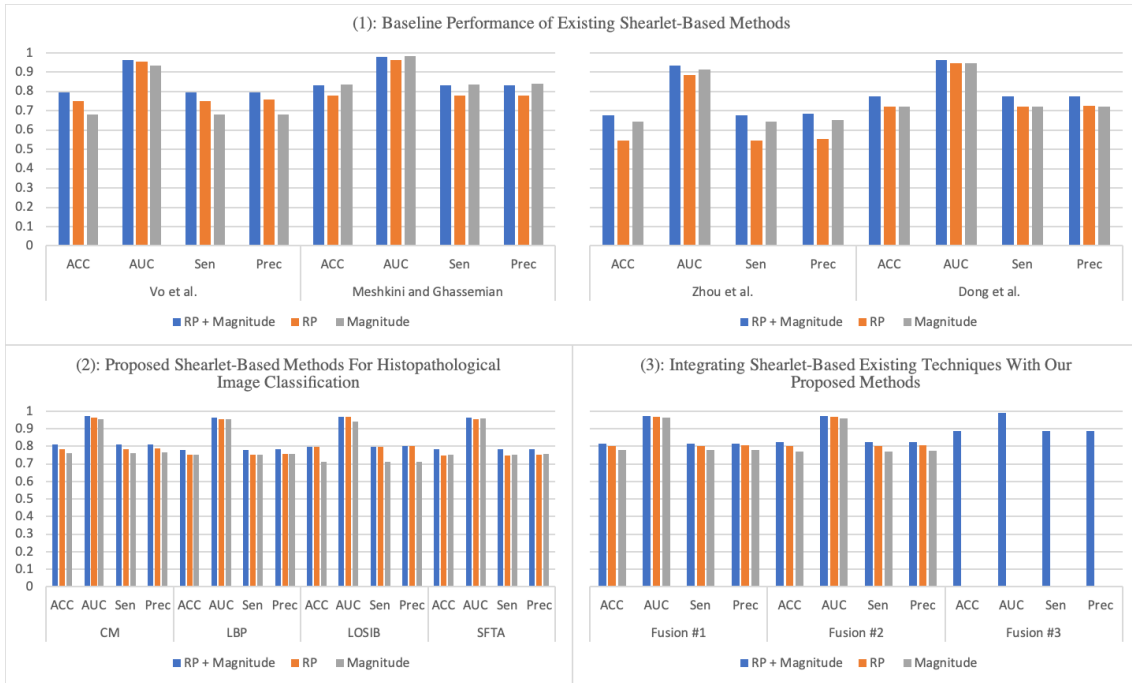
- Vo et al. [220]: The authors propose to extract the circular mean and circular variance from the RP and mean of magnitude of each sub-band. We extract the same descriptors of each shearlet sub-band and then concatenate them to form the feature vector.
- Meshkini and Ghassemian [142]: The resultant inner product of the co-occurrence matrix and magnitude coefficients of shearlet is used as a feature vector.
- Zhou et al. [241]: As proposed by the original authors, a combination of three feature types are extracted and used jointly in our experiments. The three feature types are only used on the magnitude coefficients and are computed as follows: 1) Only the first layer in the horizontal cone of shearlet is utilized to obtain the co-occurrence matrix and then the entropy, correlation, contrast and, energy are calculated; 2) From only the first and third layer coefficients in the horizontal and vertical cones of the shearlet transform, the mean, variance, and energy are computed; 3) From high-frequency sub-band coefficients, the maximal values are extracted from each column.
- Dong et al. [50]: The authors compute the mean and standard deviation of the magnitude coefficients of each sub-band to obtain the feature vector of an image.

Part (1) in Figure 3.2, 3.3, 3.4, and 3.5 presents the classification performance of the techniques as detailed above for Kather, BreakHis, Epistroma, and Warwick-QU datasets, respectively. Further, we provide the classification results in form of tables in Appendix B (i.e., the variants shown in italic in Part (1) of the Tables B.1, B.2, B.3, and B.4 are not examined in the original studies but we have added them for comparison). We notice that utilizing both of the magnitude and RP of the shearlet coefficients enhance the classification performance irrespective of the classifier (i.e., SVM or DTB) model and enhances the performance of other techniques. It can be seen that the attributes proposed by Vo et al., using both magnitude and RP, attain better performance than other baseline methods. When utilizing the multi-class Kather’s dataset, our proposed descriptors extracted from the shearlet coefficients of both magnitude and RP can achieve reasonable accuracy between 82% to 86% when using an SVM model, but DTB yields accuracy spans only between 79% to 81% as presented in Part (2) of Figure 3.2 (See Table B.1). The highest AUC value = 0.9773 is obtained while classifying Kather’s dataset utilizing LOSIB attributes coupled with the SVM model. Furthermore, we compute the Sen and Prec which the highest values are achieved using LOSIB coupled with SVM (i.e., 0.8632 & 0.8664, respectively). However,

when incorporating various descriptors together, for instance **Fusion #3** improves the accuracy about 6.22% points as presented in Part (3) of Figure 3.2 & Table B.1 when using an SVM model.



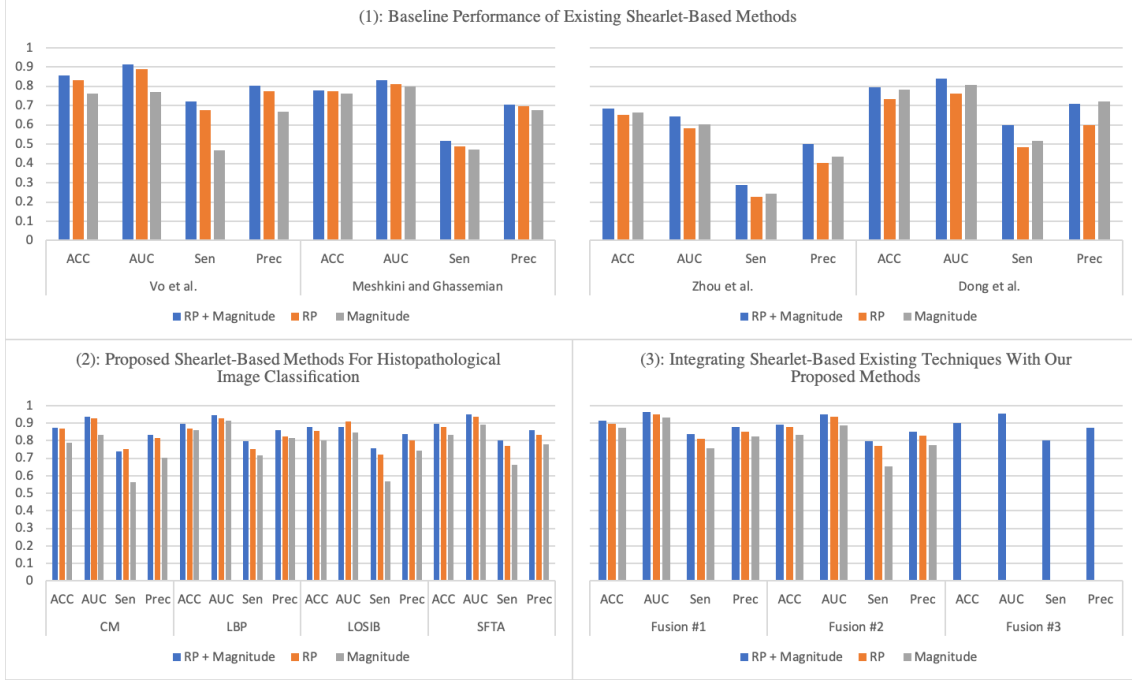
(a) SVM classification performance



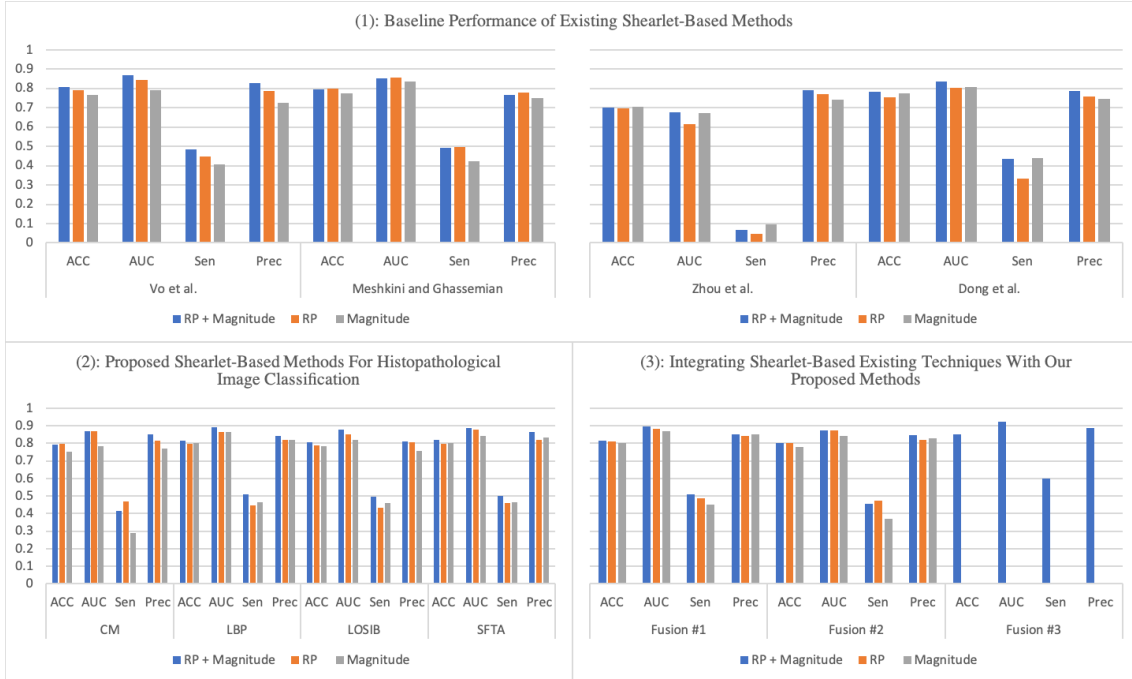
(b) DTB classification performance

Figure 3.2: Classification performance on Kather's dataset using shearlet-based descriptors.

We report the classification performance on the BreakHis dataset by considering all

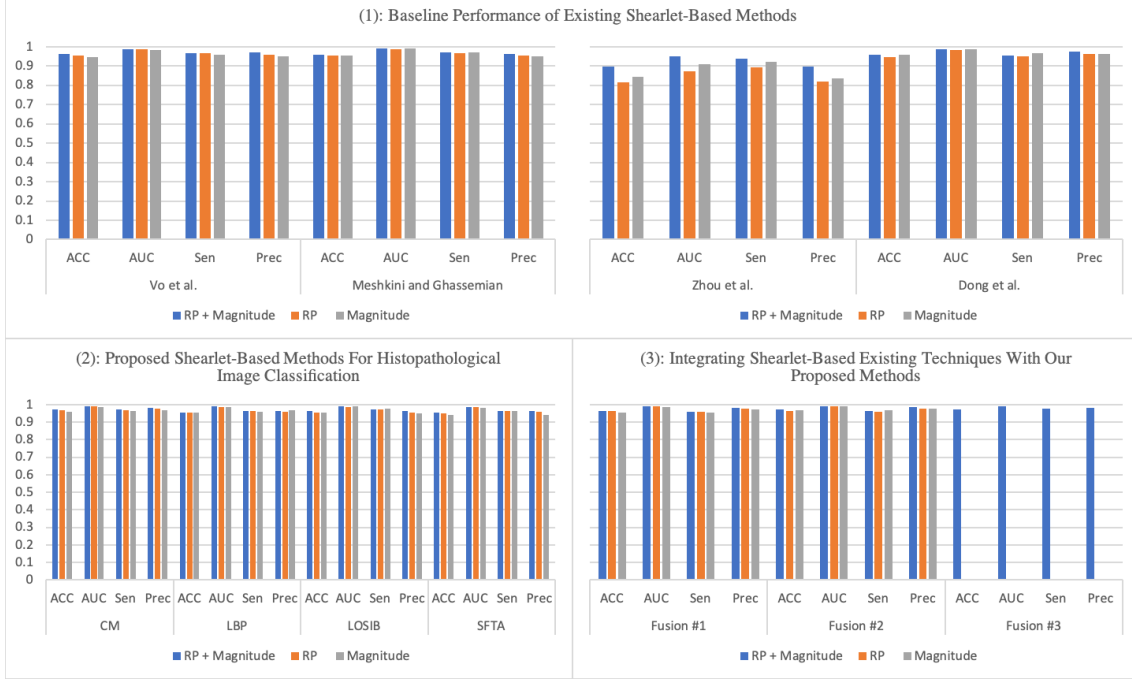


(a) SVM classification performance



(b) DTB classification performance

Figure 3.3: Classification performance on BreakHis dataset using shearlet-based descriptors.



(a) SVM classification performance



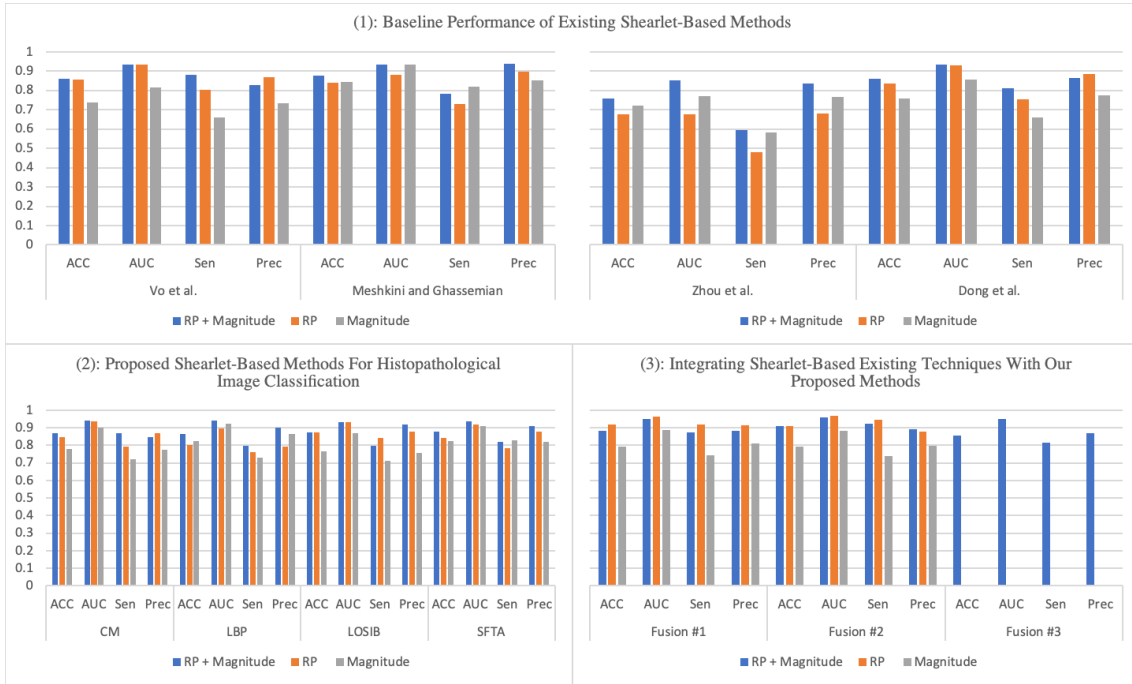
(b) DTB classification performance

Figure 3.4: Classification performance on Epistroma dataset using shearlet-based descriptors.

four magnification factors as one dataset (as seen in Figure 3.3). We observe that SFTA has attained the highest accuracy of 89.72% on the validation split (with a corresponding



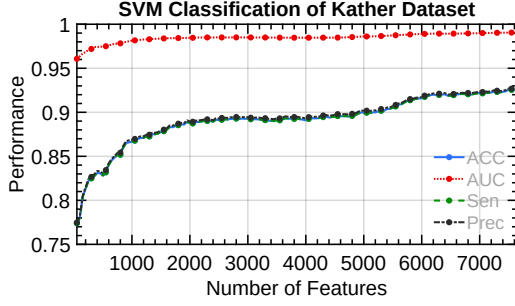
(a) SVM classification Performance



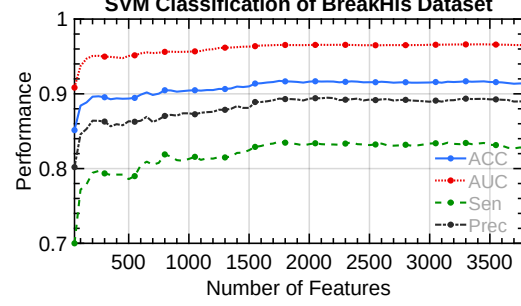
(b) DTB classification Performance

Figure 3.5: Classification performance on Warwick-QU dataset using shearlet-based descriptors.

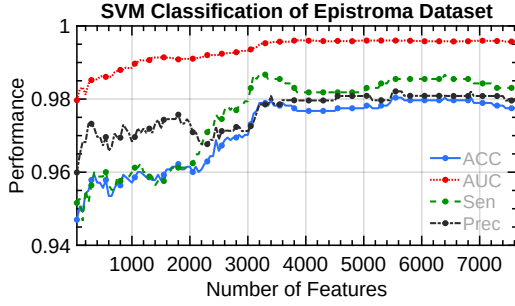
AUC = 0.9527, Sen = 0.8040, and Prec = 0.8593) while utilizing both magnitude and RP. As presented in Part (3) of Figure 3.3 (See also Table B.2), **Fusion #1** has led to



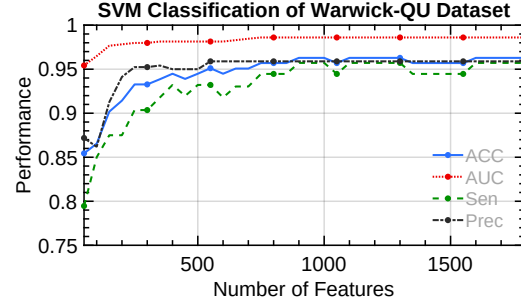
(a) Achieved ACC = 92.56% of Kather (Fusion #3) while using 7500 PCs



(b) Achieved ACC = 91.73% of BreakHis (Fusion #1) while using 1750 PCs



(c) Achieved ACC = 98.04% of Epistroma (Fusion #3) while using 5500 PCs



(d) Achieved ACC = 96.29% of Warwick-QU (Fusion #2) while using 900 PCs

Figure 3.6: Classification performance of SVM with an increment of 50 principal components (PCs) of best Fusion till all PCs are used.

a higher accuracy of 91.28% (with an AUC = 0.9650, Sen = 0.8391, and Prec = 0.8775). Evidently, due to the high skewness in the BreakHis dataset, the classifier is biased toward the majority class (i.e., malignant cases) as it is observed with lower values in sensitivity and precision.

Similarly, we conduct our technique on the Epistroma dataset (results reported in Figure 3.4 & Table B.3). We observe when descriptors extracted from both, magnitude and RP, utilizing CM leads to an accuracy of 97.24% with AUC = 0.9917, Sen = 0.9733, and Prec = 0.9807. In Part (3), **Fusion #3** enhances the accuracy to 97.46% with a AUC value = 0.9925, Sen = 0.9769, and Prec = 0.9809. Moreover, the classification performance on Warwick-QU are is presented in Figure 3.5 (See also Table B.4). Once more, CM descriptors of both, magnitude and RP lead to the best accuracy of 95.70% (with value AUC = 0.9860, Sen = 0.9446, and Prec = 0.9589) of our proposed individual feature representation. Part (3) presents that **Fusion #2** leads to the highest accuracy of 96.29% with an AUC = 0.9860, Sen = 0.9571, and Prec = 0.9589. It is worth noting that the reported standard deviations corresponding with classification performance on this dataset are higher than the other datasets because the number of validation examples of this dataset is significantly smaller. As shown in Fig. 3.6, we attempt to reduce our feature representation for each dataset. As such, we further process the combined descriptors of the best achieving fusion method with PCA. Our goal is to keep the PCs

Kather Dataset Classification Using SVM with PCA

True Class \ Predicted Class	Tumour	Stroma	Complex	Lympho	Debris	Mucosa	Adipose	Empty
Tumour	582	10	23	3	1	6		
Stroma	5	556	23	6	29	6		
Complex	28	29	528	29	5	6		
Lympho	2	3	35	580		5		
Debris	1	23	15	1	570	7	8	
Mucosa	14	2	9		5	595		
Adipose		7			7		603	8
Empty					1		7	617

(a) Kather classification (Fusion #3) confusion matrix

BreakHis Dataset Classification Using SVM with PCA

True Class \ Predicted Class	Benign	Malignant
Benign	2074	406
Malignant	248	5181

(b) BreakHis classification (Fusion #1) confusion matrix

Epistroma Dataset Classification Using SVM with PCA

True Class \ Predicted Class	Epithelium	Stroma
Epithelium	813	12
Stroma	15	536

(c) Epistroma classification (Fusion #3) confusion matrix

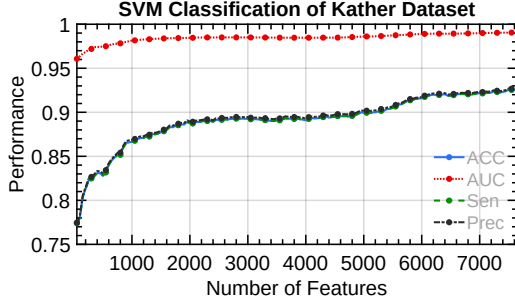
Warwick-QU Dataset Classification Using SVM with PCA

True Class \ Predicted Class	Benign	Malignant
Benign	71	3
Malignant	3	88

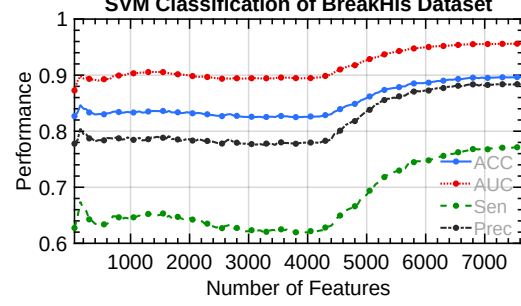
(d) Warwick-QU classification (Fusion #2) confusion Matrix

Figure 3.7: Confusion matrices of our best performing proposed methods while using reduced set of features (see Figure 3.6).

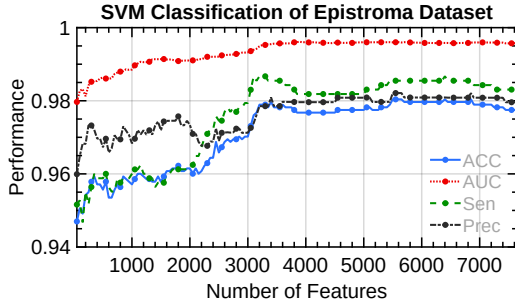
that maintain or increase the accuracy when compared to using all descriptors. We have established from Table B.1 - Table B.4 that the SVM model has strong capabilities to classify histopathological tissues which is therefore coupled with the best fusion of the corresponding dataset to find a reduced set of features. In case of the Kather dataset, we were only able to reduce the number of features of Fusion #3 from 7648 to 7500 while maintaining high accuracy of $92.56\%(\pm 1.29\%)$ with a corresponding $AUC = 0.9905(\pm 0.0020)$, $Sen = 0.9256(\pm 0.0130)$, and $Prec = 0.9270(\pm 0.0125)$. A drastic reduction of attributes from 3776 to 1750 features for the BreakHis dataset was possible while slightly improving accuracy = $91.73\%(\pm 0.59\%)$ with $AUC = 0.9654(\pm 0.0057)$, $Sen = 0.8363(\pm 0.0188)$, and $Prec = 0.8936(\pm 0.0169)$. Again, the accuracy of classifying Epistroma increases slightly to $98.04\%(\pm 1.03\%)$ with $AUC = 0.9960(\pm 0.0043)$, $Sen = 0.9867(\pm 0.0120)$, and $Prec = 0.9809(\pm 0.0136)$ when reducing the number of descriptors from 7648 to 6400. Furthermore, to classify the Warwick-QU dataset, only 900 (out of 1792 attributes) are needed to achieve the highest accuracy of $96.29\%(\pm 7.89\%)$ with $AUC = 0.9860(\pm 0.0353)$, $Sen =$



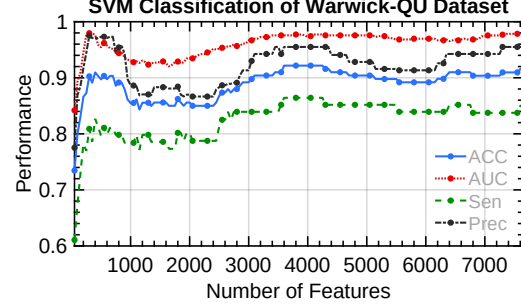
(a) Achieved ACC = 92.56% of Kather (Fusion #3) while using 7500 PCs



(b) Achieved ACC = 89.62% of BreakHis (Fusion #3) while using 7250 PCs



(c) Achieved ACC = 98.04% of Epistroma (Fusion #3) while using 5500 PCs



(d) Achieved ACC = 92.17% of Warwick-QU (Fusion #3) while using 3600 PCs

Figure 3.8: Classification performance of SVM with an increment of 50 PCs of Fusion #3 till all PCs are used.

0.9571(± 0.0964), and Prec = 0.9589(± 0.0945). For a completion for feature reduction, we provide the confusion matrices of each best performing SVM model while using the reduced set of features as shown in Figure 3.7.

Furthermore, the complexity of histopathological images differs from one dataset to another. Although our proposed individual shearlet-based descriptors are capable of achieving reasonable accuracy across the four histopathological image datasets, the best fusion which can be used to achieve the highest potential accuracy classifying the histological image dataset is not always the same (as shown in Fig. 3.6). Considering Fusion #3 of our shearlet-based descriptors expressed in the principal components, we can observe different patterns of classification performance while using an increment of 50 PCs in Fig. 3.8. Therefore, our shearlet-based descriptors appear to be general enough for any histological image analysis, and carefully choosing a type of fusion that maximizes the recognition of the tissue type is sufficient. In the literature exist state-of-the-art techniques that use the same datasets that we have used in our study. We report their results in this section to show that our proposed techniques are on par with those techniques and achieve excellent performance in terms of accuracy and AUC. Our proposed approach achieves competitive results in all cases making it an appealing technique for pattern recognition in medical imaging. Wang et al. [222] achieve an accuracy of 92.6% (and AUC = 0.985) while training/validating a bilinear CNN with Kather’s dataset, but our proposed approach achieves

comparable accuracy, yet better AUC as shown in Table B.1. Furthermore, Jonnalagedda et al. [101] were able to achieve an accuracy of 92.2% (and AUC = 0.92) while using a deep learning model (with augmented images and information from tumor nuclei) on BreakHis; evidently, our approach is capable of achieving higher AUC but with slightly lower accuracy.

In the case of the Epistroma dataset, Ramalho et al. [175] achieved an accuracy of $\approx 95\%$ based on a structural approach using co-occurrence statistics; however, we achieve better accuracy. Ribeiro et al. [177] proposed to extract a set of features (i.e., Fractal measures, Haralick features, and curvelet) from Warwick-QUC for which they obtain a higher AUC of 0.994 but they did not report accuracy.

We show that our technique competes with the state-of-the-art including those based on CNN in terms of classification accuracy. Our approach is based on hand-crafted features and gives comparable results to resource-extensive CNN-based approaches. The approach based on different descriptors in our study is capable of handling the limited number of samples in biomedical datasets. The highest accuracies in our study are obtained using an SVM model which is considered as a black box and the classifications made by such models are difficult to interpret and explain [53]. In addition to SVM, we explored DTB which is capable of achieving reasonable results, and might be a more favorable approach in certain applications as they have higher interpretability [74]. However, navigating through a rule to understand a decision made by a DTB is still a challenge since our main descriptors are computed based on shearlet coefficients. Therefore, linking the computed statistical information from the shearlet transform to a particular region of interest in a tissue type can be difficult. It is worth noting that our main concern in this study is to establish a robust approach for classifying histological images.

3.7 Summary

In this chapter, we have presented our proposed technique to build an automated framework for the classification of histopathological images based on novel shearlet descriptors. Such descriptors are robust to predict tissue types. We use a complex shearlet transform where we can leverage the multi-resolution and multi-directionality to characterize a tissue type. Further, we can compute phase and magnitude coefficients from the complex coefficients. Therefore, we have computed various local and global statistics that encode and summarize the phase and magnitude coefficient information. Our hand-crafted approach was effective in achieving competing classification results while handling small size histopathological datasets which present different levels of complexity.

Chapter 4

Histological Image Classification using Deep Features and Transfer Learning

4.1 Overview

In this chapter, we study the benefits of using pre-trained deep learning models as unsupervised feature extractors and transfer learning for histopathological image classification. Given that well-annotated histopathological datasets are scarce, training supervised deep neural networks is a challenging issue [171]. Various studies reported remarkable classification results [101, 205] that are based on training a DL model either from scratch or opting to investigate a shallow DL model. Such studies, which also utilized supervised DL strategies, delivered effective models with the cost of using aggressive data augmentation and potentially long training time to reach an optimal model. Instead, we would like to study the applicability of pre-trained CNNs model while utilizing publicly available histopathological datasets that present varying levels of complexity.

Four state-of-the-art deep learning models are used as unsupervised feature extractors, as described in Section 4.3. These pre-trained models are utilized to obtain a more abstract representation, i.e., a feature vector for each histopathological tissue. Then, an SVM model is trained to classify the histopathological image, i.e., to quantify the effectiveness of such DL models to map a tissue. Furthermore, the same four models are re-trained end-to-end without any data augmentation, as described in Section 4.4. Finally, two of the smaller models, i.e., SqueezeNet-v1.1 and ResNet-18, are selected to examine the effect of block-wise fine-tuning. The same smaller models are also used to examine the effect of varying the number of tissue samples per class in the context of colorectal cancer histology on the multi-class Kather’s dataset. The goal of this chapter is to benchmark the accuracy of deep features as compared to fine-tuning in histopathological image applications. In Section 4.6, we compare the performance of our techniques and other exiting approaches using the same datasets.

4.2 Deep Models For Histopathological Images

We use pre-trained models that are trained on large datasets of non-medical images (i.e., ImageNet). These pre-trained models are used either as feature extractors or weight initializers for histopathological image classification. We then compare the performance of 1) an SVM model trained and tested on the deep features extracted from pre-trained models (see Section 4.3); 2) Re-trained models (initialized with pre-trained weights) with minimal changes in the model architecture for different tissue types classification (see Section 4.4).

Various models are designed to capture intrinsic information from the input data in a CNN model. In this study, we examine lightweight networks, namely, SqueezeNet version 1.1 [93] and MobileNet version 2 [186]. In addition, we also would like to examine the applicability of larger networks, namely, ResNet-18 [83] and DenseNet-201 [89]. In particular, SqueezeNet exploits filters of various sizes that are potentially capturing fine-grained and coarse attributes of a histopathological image. The authors of MobileNet-v2 show that their model is faster than, e.g., ShuffleNet [237] and NASNet [244]. On the other hand, DenseNet-201 exploits previously obtained feature maps, i.e., feature reuse. As such, they are propagated and concatenated with current feature maps in a depth-based manner, which inputs more information to the subsequent layers. Unlike DenseNet-201, the ResNet-18 model concatenates feature maps, but using element-wise operation.

Other DL models of similar nature were explored in our study. For instance, Inception [209] is similar to the DenseNet design, which also uses skip connection to concatenate feature maps in a depth-based manner, but the Inception building block is wider and generated by kernels of different sizes. When Inception was examined for histopathological image classification, it was not as effective as DenseNet. Furthermore, there are other older models that did not employ the concept of skip connection, and such models can potentially be slow to train and face vanishing gradients, e.g., AlexNet [112] and VGG architectures [195]. These models were also examined for histopathological image classification and did not yield promising results in our experiments.

A brief description of the four models that are used in this study is as follows.

- SqueezeNet-v1.1 [93]: This network starts with a convolution layer (conv1), followed by eight blocks of 2-9 fire modules. A fire block starts with a squeeze convolution layer of size 1×1 which goes into two expand layers, one that has a filter size of 1×1 and the other has a size of 3×3 . The feature maps of the two feed into a concatenation layer which is an input to the squeeze layer that outputs to the following fire module.
- MobileNet-v2 [186]: This model introduces two types of blocks: 1) combination of linear bottleneck operations, and 2) skip connection. Both blocks contain the following operations: convolution, batch normalization, and modified rectified linear unit (i.e., $\min(\max(x, 0), 6)$). This model has a total of 16 blocks.
- ResNet-18 [83]: This network consists of 8 modules where each module consists of two convolution layers both have a filter size of 3×3 , and each filter is followed by

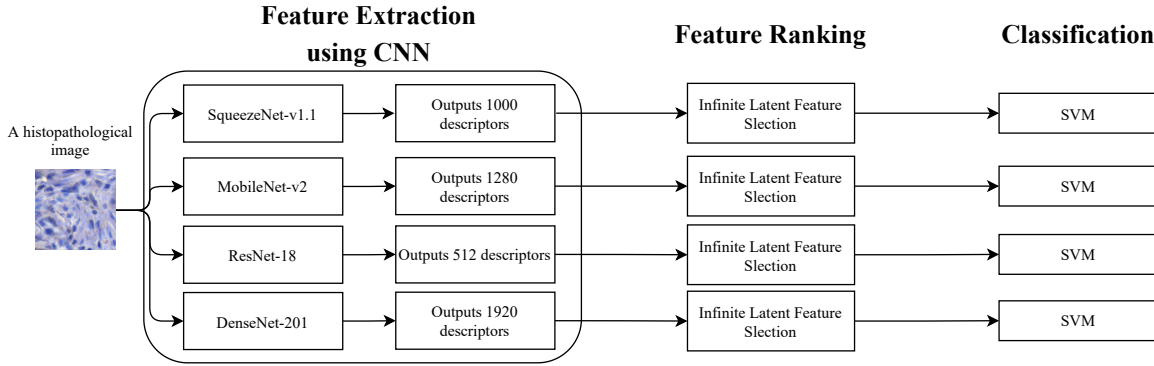


Figure 4.1: Our approach for using deep features for histopathological image classification. © Copyright IEEE 2020.

batch normalization. The Element-wise addition is applied to both, the output of a previous module with the current one.

- DenseNet-201 [89]: The architecture of this network is also based on the concept of residual learning introduced with ResNet [83]. Unlike ResNet which applies the element-wise addition of previous feature maps, DenseNet applies depth concatenation of both current and previous outputs. DenseNet-201 consists of 32-blocks. Each dense block consists of two standalone convolution layers with kernel sizes 1×1 and 3×3 , respectively, and convolution layers preceded by batch normalization.

4.3 CNN as Feature Extractor

An overview of our study of histological images using deep features is shown in Fig. 4.1. Our method is implemented using MATLAB® 2019a. The four pre-trained CNN architectures described in Section 4.2 are used as feature extractors. As such, the output of the deepest layer (i.e., the layer just before the classification layer) of a pre-trained model is flattened to obtain the feature vector of each histopathological image without any fine-tuning. Thereafter, a standard classifier, namely SVM, is trained and tested to evaluate the capability of such deep features to classify histopathological images.

4.3.1 Infinite Latent Feature Selection

As shown in Figure. 4.1, the number of features generated from the pre-trained CNN models have high-dimensionality. High dimensionality with a limited number of histopathological samples may lead to over-fitting [207]. Therefore, we propose to use Infinite Latent Feature Selection (ILFS) [179] that was shown to be effective with deep features. As opposed to other feature reduction methodologies, ILFS does not generate a new feature space, it only ranks feature importance.

Infinite Latent Feature Selection (ILFS): Dimensionality reduction using ILFS is based on probabilistic latent graph-based feature ranking. Sub-spaces of all attributes are considered as paths on a graph. As such, sub-spaces of attributes (i.e., nodes) are connected via paths in a graph. Furthermore, the importance of a feature is revealed through a conditional probability, where the latent variable is the relevance of attributes.

In the next section, we describe the procedure of using the same deep models as weight initializers for transfer learning.

4.4 Transfer Learning For Histopathological Images

Transfer learning can be categorized as transductive and inductive learning [183]. In the case of transductive transfer learning, the training dataset in both source and target domain are of similar nature. Most importantly, transductive transfer learning deals with the same classification task in both the source and target datasets. On the other hand, inductive transfer learning rather deals with different classification tasks in both the source and domain datasets.

Wei et al. [226], and Sharma and Mehra [192] adopted pre-trained weights of a CNN model. These weights are fine-tuned for the classification task of breast cancer, i.e., inductive transfer learning. They showed that fine-tuning a model for a different classification task helps in achieving better results than initializing the model with random weights, i.e., training from scratch. Similarly, we apply inductive transfer learning in our study. As such, each of the CNN architectures described in section 4.2 are fine-tuned, but for different classification tasks. Each of these models is adopted as provided in MATLAB® 2019a with simple adjustments made such that the models can be re-trained for histopathological image classification. However, we remove the head of the model, i.e., the specialized layers prior to the classification layer, which instead is trained from scratch. We examine our DL strategies on four publicly available datasets. One of them consists of eight labels while the others are binary datasets. Therefore, our SoftMax classification layer must be trained from scratch, which has either two or eight outputs.

4.4.1 CNN Fine-tuning

To retrain the CNN models, we choose stochastic gradient descent as an optimizer [145] with a momentum value of 0.9. Following the success of previous studies [105], we also choose a small learning rate of 0.0003. The intuition behind choosing such momentum and learning rate is to make sure to reduce the oscillation and slowly updating the weights adopted from the pre-trained models, respectively. Finally, the CNN models are retrained for a maximum of 20 epochs with mini-batches of size 10.

In Section 4.6, we compare the performance of the SVM models based on the various deep features computed from different CNN architectures with retraining the deep models.

Table 4.1: Summary of datasets used in Ch. 4.

DB	#Classes	#Samples	Sample Size
Kather’s	8	5000	150×150
BreakHis	2	7909	700×460
Epistroma	2	1376	Variable
Warwick-QU	2	165	Variable

4.5 Experiments

4.5.1 Datasets

Deep features and fine-tuning techniques are examined on four histopathological datasets of different complexity levels, as described in Section 2.2.1). As a reminder, we provide a summary of these four datasets in Table 4.1.

4.5.2 Cross-validation (CV)

Cross-validation (CV) is a common technique to obtain more general results about classification models. In our study, we adopt 10-fold CV to split the following three datasets: Kather’s (as Kather et al. [103]), Epistroma (as Ramalho et al. [175]), and Warwick-QU (as Ribeiro et al. [177]). In 10-fold CV, 9-folds are utilized to train a model, and 1-fold is utilized to test this model; hence, this is repeated 10 times and finally, the mean and standard deviation is reported. On the other hand, the BreakHis dataset is divided into 7-folds CV as Jonnalagedda *et al.* [101].

4.6 Results and Analysis

We quantify the performance of our method on the four datasets (See Section. 4.5.1), we use common metrics for classification: **accuracy** and **AUC** (area under receiver operating characteristic curve). The summary of our results of using the full set of deep features to train an SVM model is shown in Table. 4.2. We apply ILFS to rank features and use an increment of 50 descriptors at a time to train an SVM model. We report the results with sets of descriptors that result in an accuracy at least as high as the accuracy of using the full set of features (see Table. 4.3). Table. 4.4 shows the results of fine-tuning all layers of various CNN models. Table 4.5 and Table 4.6 shows the classification performance of block-wise fine-tuning of both SqueezeNet-v1.1 and ResNet-18, respectively. We provide a comparison with state-of-the-art techniques showing the robustness of our methods in Section 4.6.3.

Table 4.2: SVM classification performance utilizing deep features of various pre-trained models.

Model		Kather	Epistroma	Warwick	BreakHis
Squeeze-Net	ACC%	91.22% $\pm$ 0.0117%	97.17% $\pm$ 1.35%	90.92% $\pm$ 7.17%	92.53% $\pm$ 0.63%
	AUC	0.9886 $\pm$ 0.0022	0.9958 $\pm$ 0.0045	0.9413 $\pm$ 0.0741	0.9697 $\pm$ 0.0055
Mobile-Net	ACC%	91.48% $\pm$ 1.18%	97.96% $\pm$ 0.89%	98.24% $\pm$ 2.84%	90.19% $\pm$ 0.96%
	AUC	0.9875 $\pm$ 0.0013	0.9976 $\pm$ 0.0039	1 $\pm$ 0	0.9596 $\pm$ 0.0058
ResNet	ACC%	92.80% $\pm$ 0.75%	98.55% $\pm$ 1.14%	98.79% $\pm$ 2.56%	94.61% $\pm$ 0.66%
	AUC	0.9912 $\pm$ 0.0009	0.9968 $\pm$ 0.0042	1 $\pm$ 0	0.9847 $\pm$ 0.0039
Dense-Net	ACC%	95.22% $\pm$ 0.57%	99.20% $\pm$ 0.87%	100% $\pm$ 0%	97.90% $\pm$ 0.52%
	AUC	0.9957 $\pm$ 0.0006	0.9984 $\pm$ 0.0038	1 $\pm$ 0	0.9974 $\pm$ 0.0014

4.6.1 Classification Results Using Kather’s dataset

First, we report results using the selected CNN models as unsupervised feature extractors (see Section 4.3). Images must be reshaped to fit the input layer of the pre-trained CNN model. After that, the feature vectors extracted from a CNN model are fed to an SVM model. Although the pre-trained CNN models were trained on the ImageNet dataset, we observe their capability of being effective on histopathological images. For instance, DenseNet-201 classified the 8-class Kather dataset with a reduced set of descriptors (1000 instead of 1920 descriptors) with an accuracy of 95.40% (with corresponding AUC value = 0.9961). However, fine-tuning the simpler models results in about 3% point improvement compared to using the same model as a feature extractor.

Next, we report our results for fine-tuning a CNN model for Kather’s dataset. We adjust the input layer to fit the dimensions of the data (i.e., 150×150 pixels). By fully fine-tuning DenseNet-201, we achieve an accuracy of 95.02% with an AUC value = 0.9976 (see Table 4.4). However, DenseNet-201 is our largest model and smaller models (e.g., SqueezeNet) may be preferable as they require significantly less computational power. We found overall performance quite comparable with an accuracy of 94.18% (with AUC value = 0.9964). Additionally, we select two models (i.e., SqueezeNet and ResNet) to examine the appropriate level of blocks of a model to be fine-tuned. The fewer blocks need to be fine-tuned, the lesser is the computational demands for training. As described in Section 4.4, both of the selected models consist of 8-modules, unlike MobileNet and DenseNet which consists of 16 and 32 modules, respectively. We observe that only the last 5 modules of SqueezeNet are necessary to be retrained to achieve an accuracy of 93.24% (with AUC value = 0.9948). Fine-tuning the shallower blocks does not change the accuracy of the model by much. Similarly, we observe only the 4 deepest blocks of ResNet-18 need to be fine-tuned to obtain an accuracy of 93.14% (with AUC value = 0.9948). Therefore, fine-tuning the appropriate level of blocks will significantly accelerate the retraining process of a CNN model while maintaining high accuracy. Being able to determine the appropriate

Table 4.3: SVM classification performance utilizing top ranked deep features of various pre-trained models.

Model		Kather (#Features)	Epistroma (#Features)	Warwick (#Features)	BreakHis (#Features)
Squeeze-Net	ACC%	91.02% $\pm$ 1.08%(650)	97.67% $\pm$ 1.02%(300)	90.33% $\pm$ 7.64%(550)	92.54% $\pm$ 0.59%(900)
	AUC	0.9884 $\pm$ 0.0023(650)	0.9952 $\pm$ 0.0047(300)	0.9343 $\pm$ 0.0727(550)	0.9695 $\pm$ 0.0058(900)
Mobile-Net	ACC%	91.30% $\pm$ 1.08%(900)	98.26% $\pm$ 1.14%(300)	98.82% $\pm$ 2.48%(850)	90.18% $\pm$ 0.91%(1100)
	AUC	0.9874 $\pm$ 0.0013(900)	0.9977 $\pm$ 0.0032(300)	1 $\pm$ 0(850)	0.9575 $\pm$ 0.0060(1100)
ResNet	ACC%	92.72% $\pm$ 1.01%(250)	98.55% $\pm$ 1.14%(200)	100% $\pm$ 0%(150)	94.30% $\pm$ 0.57%(350)
	AUC	0.9910 $\pm$ 0.0013(250)	0.9977 $\pm$ 0.0025(200)	1 $\pm$ 0(150)	0.9812 $\pm$ 0.0044(350)
Dense-Net	ACC%	95.40% $\pm$ 0.55%(1000)	99.06% $\pm$ 0.84%(250)	100% $\pm$ 0%(200)	97.96% $\pm$ 0.57%(1300)
	AUC	0.9961 $\pm$ 0.0006(1000)	0.9978 $\pm$ 0.0042(250)	1 $\pm$ 0(200)	0.9974 $\pm$ 0.0014(1300)

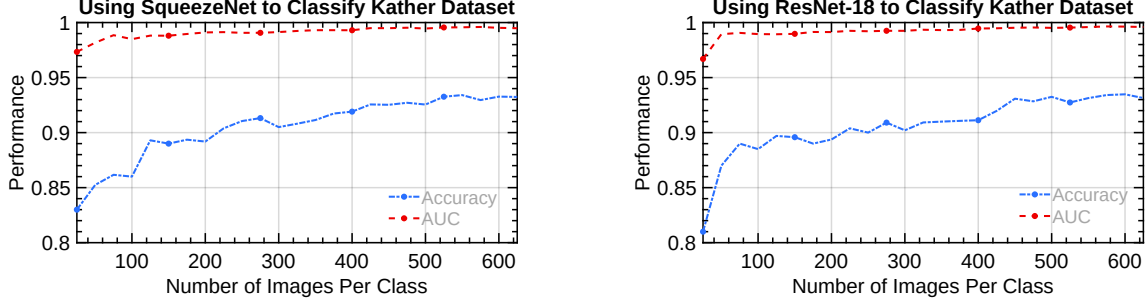
level of blocks to fine-tune while using all images per class of Kather dataset, we again fine-tune the appropriate level of blocks for both SqueezeNet and ResNet but this time varying the number of images used for fine-tuning. We vary the number of images in increments of 25 images per class until all 625 images per class are used as shown in Figure 4.2. We achieve very low classification performance, i.e., the accuracy of 83.00% and 81.00%, for SqueezeNet and ResNet, respectively when only 25 images are used. However, both networks are trainable with less than all images and SqueezeNet achieves an accuracy of 93.41%, and ResNet achieves 93.14% utilizing 550 images per class. In this case study of colorectal cancer histology, we need 550 images per class to achieve accuracy close to using all images but even with much fewer images, reasonable accuracy can be obtained (see Figure 4.2).

4.6.2 Classification Results Using Epistroma, Warwick-QU, and BreakHis datasets

For further validations, we apply our approach on three additional varied datasets. We report results on the BreakHis dataset considering all magnification factors as one dataset. Similarly, we apply our approach to the Epistroma and Warwick-QU datasets.

We observe a similar behavior using DenseNet-based descriptors with the three additional datasets. They provide the best results when fed to an SVM model:

- Epistroma: Only 250 descriptors extracted from DenseNet are needed to achieve an



(a) Fine-tuning Fire5 block and after of SqueezeNet (b) Fine-tuning Res4a block and after of ResNet-18

Figure 4.2: Varying the number of images per class of multi-class Kather’s dataset (8-class dataset). Starting with 25 images per class till all available images are used. © Copyright IEEE 2020.

accuracy of 99.06% (with an AUC value = 0.9978). However, when all descriptors of DenseNet (i.e. 1920) are used, we observe a slight enhancement of accuracy = 99.20% (with AUC value = 0.9984).

- Warwick-QU: On this dataset, ResNet-18 and DenseNet-201 both achieve perfect accuracy/AUC. However, ResNet-18 has an edge over DenseNet since only 150 descriptors compared to 250 descriptors are needed to train an SVM model for perfect accuracy.
- BreakHis: Only 1300 out of 1920 DenseNet descriptors are needed to achieve an accuracy of 97.96% (with corresponding AUC value = 0.9974).

In the case of fine-tuning, the input layer is replaced to fit the dataset dimensions. BreakHis images have a uniform size of 700×460 , unlike Epistroma and Warwick which consist of variable size images; hence, the images are resized to 172×172 and 430×430 pixels, respectively.

Each of the additional three datasets being used in our study (i.e., BreakHis, Epistroma, and Warwick-QU) are also utilized to retrain and test the four CNN models. We also examine block-wise fine-tuning of SqueezeNet and ResNet on the three varied datasets as shown in Table 4.5 and Table 4.6, respectively. Again, we observe the whole CNN model does not need to be fully fine-tuned. However, the appropriate level of blocks to be fine-tuned is dependent on the complexity and the number of samples of the histopathological dataset. Although the datasets are limited in the number of samples, our strategy of fine-tuning the CNN models provides very competitive results.

4.6.3 State-of-the-art Results

Many studies have proposed to identify the tissue types from the pattern revealed in the hematoxylin and eosin (H & E) stained slides on the tumor samples. In this section, we

Table 4.4: Classification performance utilizing pre-trained models as weight initialization, CNNs are fully fine-tuned.

Model		Kather	Epistroma	Warwick	BreakHis
Squeeze-eNet	ACC%	94.18% $\pm$ 1.49%	98.11% $\pm$ 1.04%	97.54% $\pm$ 4.34%	96.54% $\pm$ 1.29%
	AUC	0.9964 $\pm$ 0.0016	0.9955 $\pm$ 0.0053	0.9984 $\pm$ 0.0050	0.9947 $\pm$ 0.0024
Mobile-Net	ACC%	93.78% $\pm$ 1.40%	98.40% $\pm$ 0.75%	97.54% $\pm$ 4.34%	96.16% $\pm$ 1.00%
	AUC	0.9958 $\pm$ 0.0019	0.9973 $\pm$ 0.0032	1 $\pm$ 0	0.9934 $\pm$ 0.0028
ResNet	ACC%	94.14% $\pm$ 0.81%	98.18% $\pm$ 0.85%	98.20% $\pm$ 2.90%	97.57% $\pm$ 0.41%
	AUC	0.9962 $\pm$ 0.0017	0.9948 $\pm$ 0.0059	1 $\pm$ 0	0.9973 $\pm$ 0.0006
Dense-Net	ACC%	95.02% $\pm$ 1.16%	98.40% $\pm$ 0.67%	96.95% $\pm$ 3.22%	97.40% $\pm$ 0.74%
	AUC	0.9976 $\pm$ 0.0011	0.9956 $\pm$ 0.0036	0.9986 $\pm$ 0.0044	0.9973 $\pm$ 0.0016

report the state-of-the-art results. Our techniques on the four datasets achieve the highest classification performance either using deep features or fine-tuning a CNN model.

When using Kather’s dataset, Kather et al. [103] proposed using a fusion of texture-based features with a reported accuracy of 87.4% (with AUC value = 0.976). On the other hand, the state-of-the-art result of classifying Kather’s dataset is 93.4% which is achieved by Bianconi et al. [22]. They use a modified version of the local binary pattern (LBP) that consists of inter and intra descriptors with the fusion of local color contrast. Wang et al. [222] proposed to train a bilinear-CNN using 5-fold cross-validation, achieving an accuracy of 92.6% (with AUC = 0.985, sensitivity of 92.8%, and specificity of 98.9%). However, Cascianelli et al. [32], studied different deep features and achieved an accuracy of 84.8% using fast VGG model descriptors reduced using PCA.

For classification in the Epistroma dataset, Nava et al. [154], Ramalho et al. [175], and Linder et al. [126] proposed statistical and texture-based techniques. However, Bianconi et al. [22] reported the highest accuracy of 97.3% classifying the Epistroma dataset while training and testing a parameter-free nearest neighbor classifier. However, Bianconi et al. do not report AUC.

When using the Warwick-QU dataset, Ribeiro et al. [177] report the highest AUC value of 0.994 while using fractal and Haralick descriptors which are extracted both from spatial domain images and curvelet coefficients. These descriptors are fed to a polynomial (PL) classifier, trained, and tested using 10-fold cross-validation.

When using the BreakHis dataset, we compare to approaches that are training and testing a model using the different magnifications of BreakHis as one dataset. State-of-the-art results with an accuracy of 92.60% (with AUC value = 0.9565) are reported by Sharma and Mehra [192] using a fine-tuned pre-trained VGG16 model for feature extraction. Sharma

Table 4.5: Block-wise fine-tuning of pre-trained SqueezeNet for histopathological image classification.

Fine-tuned Block		Kather	Epistroma	Warwick	BreakHis
Fire9 and After	ACC%	90.22% $\pm$ 1.95%	97.67% $\pm$ 1.27%	96.36% $\pm$ 3.14%	93.45% $\pm$ 0.79%
	AUC	0.9914 $\pm$ 0.0024	0.9950 $\pm$ 0.0060	0.9954 $\pm$ 0.0105	0.9835 $\pm$ 0.0037
Fire8 and After	ACC%	91.48% $\pm$ 1.00%	97.60% $\pm$ 1.81%	96.95% $\pm$ 3.22%	94.61% $\pm$ 0.82%
	AUC	0.9927 $\pm$ 0.0020	0.9957 $\pm$ 0.0052	0.9984 $\pm$ 0.0050	0.9873 $\pm$ 0.0034
Fire7 and After	ACC%	92.18% $\pm$ 1.42%	98.40% $\pm$ 1.17%	97.57% $\pm$ 3.13%	95.42% $\pm$ 1.17%
	AUC	0.9938 $\pm$ 0.0019	0.9960 $\pm$ 0.0046	0.9984 $\pm$ 0.0050	0.9899 $\pm$ 0.0030
Fire6 and After	ACC%	92.62% $\pm$ 1.44%	98.11% $\pm$ 1.19%	96.32% $\pm$ 4.33%	96.20% $\pm$ 0.50%
	AUC	0.9937 $\pm$ 0.0029	0.9961 $\pm$ 0.0047	0.9984 $\pm$ 0.0050	0.9919 $\pm$ 0.0014
Fire5 and After	ACC%	93.24% $\pm$ 1.00%	98.18% $\pm$ 1.09%	96.95% $\pm$ 4.36%	96.34% $\pm$ 0.82%
	AUC	0.9948 $\pm$ 0.0015	0.9961 $\pm$ 0.0049	0.9984 $\pm$ 0.0050	0.9933 $\pm$ 0.0017
Fire4 and After	ACC%	93.04% $\pm$ 1.09%	98.26% $\pm$ 0.85%	98.20% $\pm$ 2.90%	96.40% $\pm$ 0.83%
	AUC	0.9953 $\pm$ 0.0016	0.9957 $\pm$ 0.0053	0.9968 $\pm$ 0.0100	0.9931 $\pm$ 0.0032
Fire3 and After	ACC%	93.30% $\pm$ 1.17%	98.18% $\pm$ 1.24%	98.79% $\pm$ 2.56%	96.82% $\pm$ 1.03%
	AUC	0.9951 $\pm$ 0.0018	0.9959 $\pm$ 0.0050	0.9952 $\pm$ 0.0151	0.9948 $\pm$ 0.0015

and Mehra, down-sampled the majority class (i.e, malignant) to equal the minority class (i.e., benign), where the balanced dataset is split into 90% for training and 10% for testing a Logistic regression (LR) classifier.

Finally, it is worth noting that we have not applied any type of data augmentation to increase the number of histological images. Data augmentation is typically used to overcome the problem of over-fitting and scarce datasets [194]. Nevertheless, data augmentation may not be ideal since many samples will be similar to each other, and performance on new data is unclear. Furthermore, classification results could be substantially worse when a CNN model is trained from scratch [226]. Therefore, we opted for adopting pre-trained models. We use a very small learning rate to slowly adjust the weights for the task we are tackling, yet bearing in mind the delivery of a re-trained model that achieves high accuracy on the testing dataset.

Table 4.6: Block-wise fine-tuning of pre-trained ResNet-18 for histopathological image classification.

Fine-tuned Block		Kather	Epistroma	Warwick	BreakHis
Res5b and After	ACC%	91.20% $\pm$ 1.50%	97.75% $\pm$ 1.06%	100% $\pm$ 0%	95.19% $\pm$ 0.67%
	AUC	0.9936 $\pm$ 0.0013	0.9953 $\pm$ 0.0042	1 $\pm$ 0	0.9915 $\pm$ 0.0022
Res5a and After	ACC%	92.16% $\pm$ 1.17%	98.25% $\pm$ 1.34%	97.57% $\pm$ 3.13%	96.55% $\pm$ 0.83%
	AUC	0.9950 $\pm$ 0.0015	0.9960 $\pm$ 0.0043	1 $\pm$ 0	0.9954 $\pm$ 0.0012
Res4b and After	ACC%	92.94% $\pm$ 1.27%	98.55% $\pm$ 0.91%	97.57% $\pm$ 4.19%	97.03% $\pm$ 1.07%
	AUC	0.9954 $\pm$ 0.0011	0.9962 $\pm$ 0.0046	1 $\pm$ 0	0.9961 $\pm$ 0.0015
Res4a and After	ACC%	93.14% $\pm$ 1.39%	98.84% $\pm$ 0.78%	99.38% $\pm$ 1.98%	96.83% $\pm$ 1.56%
	AUC	0.9959 $\pm$ 0.0011	0.9962 $\pm$ 0.0049	1 $\pm$ 0	0.9965 $\pm$ 0.0014
Res3b and After	ACC%	93.00% $\pm$ 1.24%	98.69% $\pm$ 1.01%	97.57% $\pm$ 3.13%	97.56% $\pm$ 0.76%
	AUC	0.9960 $\pm$ 0.0015	0.9964 $\pm$ 0.0045	1 $\pm$ 0	0.9973 $\pm$ 0.0014
Res3a and After	ACC%	93.36% $\pm$ 1.18%	98.76% $\pm$ 0.97%	95.22 $\pm$ 7.25%	96.77% $\pm$ 1.85%
	AUC	0.9958 $\pm$ 0.0008	0.9961 $\pm$ 0.0054	0.9984 $\pm$ 0.0050	0.9970 $\pm$ 0.0014
Res2b and After	ACC%	93.42% $\pm$ 1.17%	98.55% $\pm$ 1.28%	96.43% $\pm$ 7.45%	96.37% $\pm$ 2.98%
	AUC	0.9956 $\pm$ 0.0017	0.9962 $\pm$ 0.0054	0.9984 $\pm$ 0.0050	0.9955 $\pm$ 0.0035

4.7 Summary

In this chapter, we have studied the applicability of CNN models to small size histopathological datasets. We conclude our deep learning approach is, in fact, capable of handling the levels of complexity presented in the histopathological tissues despite the small dataset sizes. As such, the deep learning strategies can classify those small datasets with classification performance that compares favorably to methods in the literature using the same dataset. Using a DL model as an unsupervised feature extractor is a desirable strategy as minimal effort is needed to examine the effectiveness of a classifier. Furthermore, inductive transfer learning provides robust performance to classify histopathological images. It is worth noting that shallower DL models, e.g., SqueezeNet, benefit more from fine-tuning. SqueezeNet and ResNet are studied for the appropriate level of fine-tuning. Comparable results to fine-tuning all layers are obtained by fine-tuning only a few layers.

Chapter 5

Classification of Alzheimer’s Disease based on MRI Data

5.1 Overview

Conventional and deep learning techniques are proposed for AD diagnosis [214, 56]. In Chapters 3 and 4, we have proposed novel techniques to differentiate various tissue types. Our techniques based on hand-crafted features are capable of handling small size histopathological datasets that present different levels of complexity. Furthermore, we have developed deep learning strategies to classify those small datasets with surpassing classification performance in comparison to those available in the literature. In this chapter, we investigate whether our previous techniques can be extended to challenging data classification in the field of radiology, namely magnetic resonance imaging (MRI).

However, magnetic resonance imaging presents an additional challenge, namely the curse-of-dimensionality. An MRI sample of a patient brain consists of a large number of slices. As a result, the construction of a feature vector of a patient sample will contain a large number of descriptors if the same techniques as for histopathological images are used. When the number of descriptors is very high and the number of samples is very low, we are faced with the challenge of over-fitting. Moreover, MRI slices exhibit different variability patterns. Such MRI samples are limited in resolution and contrast [13]. Further, it is challenging with MRI data to localize the exact neuroanatomical structures and recognize histological characteristics, e.g., cellular and tissue properties [13]. Therefore, with the absence of such important attributes, MRI datasets can be difficult to classify.

When we applied our shearlet-based techniques from Chapter 3 directly to MRI data, the classification results for AD were not promising. One possible reason is that Alzheimer’s data contains a large number of slices, but some slices might contain irrelevant content. As a result, when computing statistical information from each slice of an MRI, a large number of descriptors will be generated, including some of which contain irrelevant attributes. Furthermore, with a low number of training samples, it can lead to over-fitting and a less effective classification model [147]. Therefore, the goal of this study to refine our shearlet-based techniques for the classification of Alzheimer’s disease patients. To achieve this goal,

we utilize two data repositories: the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database, and the Open Access Series of Imaging Studies (OASIS).

There is no universal approach to classify all types of medical images. We are able to adjust, both, hand-crafted and deep learning strategies for the application to medical images that are fundamentally different than those explored in Chapter 3 and 4. In particular, we will report the classification of MRI samples for AD after the following changes are incorporated:

- We adjust our shearlet-based descriptors from Chapter 3. We instead use the 3D shearlet transform since we are dealing with volumetric data in this chapter, namely MRI data (See Section 5.2.1). To overcome the high-dimensionality in the 3D shearlet domain, all directionalities of a slice are averaged. Finally, a particular texture-based method, i.e., CM, LBP, LOSIB, or SFTA, is used to extract descriptors from each averaged coefficient for each slice. We concatenate all descriptors from all slices to represent an MRI sample.
- Our starting point in the context of deep learning is a CNN which is pre-trained on non-medical images in ImageNet. This pre-trained network is used for weight initialization for the high-dimensional MRI sample, i.e., each sample consists of a large number of slices. Unlike most studies in the literature, we do not train our model from scratch. In addition, many studies in the literature utilize each 2D slice as an independent example. As such, if these slices are randomly split, then some slices of the same patients can appear in the testing and training split. As a result, this will cause information leak and over-fitting leading to the report of over-optimistic results on the testing data [233].

Instead, we adopt a 2D stacked MRI slices approach to be able to fine-tune a CNN model. Therefore, a CNN model needs to be adjusted, particularly, a new input layer to fit the MRI data dimensions. As an advantage of using a 2D convolutional neural network, we can use existing pre-trained weights as weight initializers for re-training a model. Also, 3D CNN models are computationally expensive and generating significantly more intermediate outcomes during the execution of a volumetric input [131], e.g., video or MRI data, which substantial computational resources are required as opposed to using a 2D CNN model. Nevertheless, a 2D CNN model lacks learning feature maps equally within all three dimensions.

We carry out the following experiments for Alzheimer’s disease classification from MRI data: Four state of the art deep learning models are fine-tuned: SqueezeNet-v1.1 [93], MobileNet-v2 [186], Inception-v3 [209], and Xception [38]. Then, these fine-tuned models are leveraged for deep feature extraction, as described in Section 5.2.2. It has been shown that using an SVM model can lead to significant improvement in the classification results [213, 173]. Therefore, instead of using the classification layer of a deep learning model, we use two other classifiers: Support Vector Machine (SVM) [79] and Decision Tree Bagger (DTB) [26]. Finally, our major contribution in this work is a robust classification model by combining both shearlet-based descriptors and deep features.

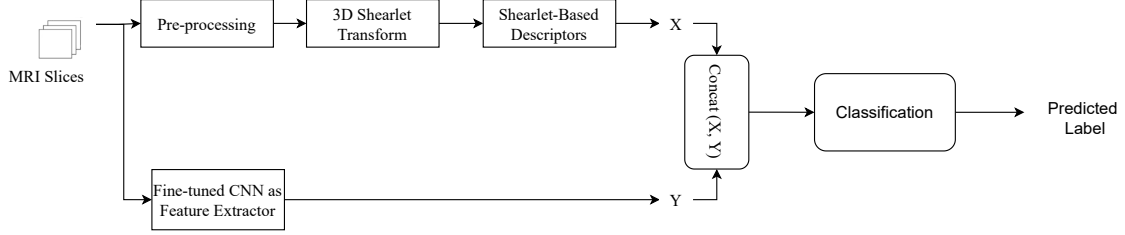


Figure 5.1: Our proposed system for integrating the 3D shearlet-based textural descriptors with deep features for the ADNI and OASIS MRI classification.

5.2 Proposed System and Components Configurations

Our proposed framework is presented in Figure 5.1. The major contribution of our work is to fuse shearlet-based descriptors with deep features. As illustrated in the figure, the proposed system consists of two pipelines. In the first pipeline of our work, an MRI sample is first pre-processed and then transformed using the 3D shearlet, and after that, the coefficients of the shearlet transform are summarized using four different texture methods: CM, LBP, LOSIB, or SFTA. Each set of descriptors are used to train either an SVM or DTB classifier, individually and jointly. The principal procedure of each summarization method differs in generating a meaningful set of descriptors for an MRI sample, and that can influence the performance of the classifier. In the case of the ADNI dataset, SFTA descriptors led to the highest classification performance. However, co-occurrence matrix textural features led to the highest classification performance for the OASIS dataset.

In the second pipeline, a variable number of CNN models are fine-tuned based on the MRI samples, and then those fine-tuned models are used as feature extractors. A previous work by Rajinikanth et al. [173] demonstrated that using an SVM classifier rather than a SoftMax layer can boost the classification performance. Therefore, in our study, instead of using SoftMax, each set of deep features is fed to another classifier either SVM or DTB. Our descriptors are used individually and jointly to train a classical ML classifier. However, we observe that MobileNet deep features that are extracted from both the ADNI and OASIS datasets achieved the highest classification performance, as we empirically show in Section 5.4.

Finally, we investigate to concatenate our proposed descriptors in two different ways: 1) shearlet-based and deep features are fed to a classical classifier, or 2) the combined shearlet-based descriptors and deep features are fed to our custom CNN model. Each component of our system is described next.

5.2.1 Shearlet-based Descriptors for AD Classification

The 3D shearlet transform [115, 116] is an extension of the 2D shearlet transform that has been discussed in Chapter 3. Hence, shearlet coefficients are also obtained using parabolic

scaling, shearing, and translation forming an affine system as in the 2D domain. Similar to the 2D domain, a ‘cone-adapted variant’ in the 3D domain is used. This variant is called pyramid-adapted shearlet systems. Such systems treat directions in a uniform scheme as in the cone-adapted in 2D shearlet transform. These pyramid-adapted discrete shearlet systems can be generated by the aforementioned affine system, in the following manner:

1. First, the systems are generated as stated by paraboloidal scaling matrices, where $j \in \mathbb{Z}$:

$$A_{2j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix}, \tilde{A}_{2j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^j & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix}, \check{A}_{2j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^j \end{pmatrix}$$

2. The directionality is captured by the shear matrices, where $k = (k_1, k_2) \in \mathbb{Z}^2$

$$B_k = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tilde{B}_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_1 & k_2 & 1 \end{pmatrix}, \check{B}_k = \begin{pmatrix} 1 & 0 & 0 \\ k_1 & 1 & k_2 \\ 0 & 0 & 1 \end{pmatrix}$$

3. The translation lattices are obtained using the following matrices:

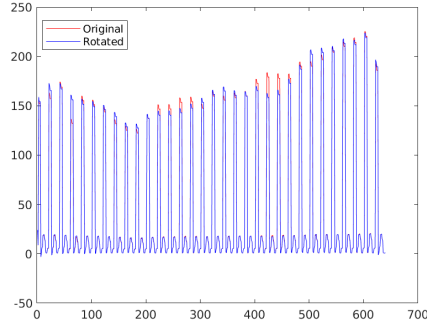
$$T_t = \text{diag}(t_1, t_2, t_2), \tilde{T}_t = \text{diag}(t_2, t_1, t_2), \text{ and } \check{T}_t = \text{diag}(t_2, t_2, t_1), \text{ where } t_1, t_2 > 0.$$

Comparable to the cone-adapted 2D shearlet transform, the 3D shearlet transform partitions the Fourier domain into a rectangular central area and six pyramids. Such partitioning in the frequency domain will generate well-localized directional filter banks that tile the complete frequency space. Therefore, those well-localized waveforms raise the ability of the 3D shearlet to handle the complex geometric discontinuities and potentially leading to generate enhanced descriptors for MRI data classification. For a further description of the shearlet transform, readers are referred to [115, 116].

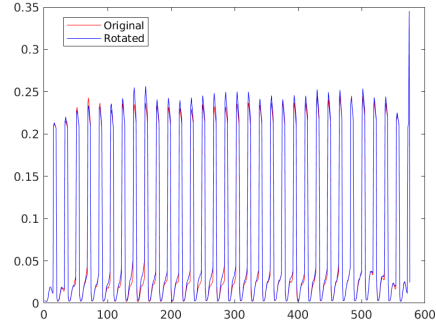
3D Shearlet Application to MRI Data

We use Shearlab [115] in our application of 3D shearlets. An MRI sample of size $h \times w \times c$ is decomposed with S scales and K orientations per scale. The dimension of the three-dimensional filters is $h \times w \times z$, which is used for convolution with the input MRI sample. This setting generates a structure of size $h \times w \times c \times z$, where h , w , c , and z denote the height, width, number of slices, and number of filters, respectively.

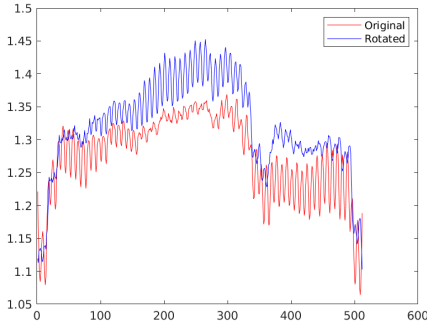
Having the magnitude coefficients of each slice with various orientations, we obtain the average presence of all shearlets; hence, the i^{th} slice is represented only by one averaged shearlet sub-band (i.e., $h \times w \times i \times 1$). Such an aggressive reduction of shearlet coefficients is necessary; otherwise, one ends up with a very large number of descriptors per slice while computing textural features from every directional sub-band. Concatenating all descriptors of all slices to form the feature vector of an MRI sample will lead to a very large number of descriptors. With a very large number of descriptors and a low number of training



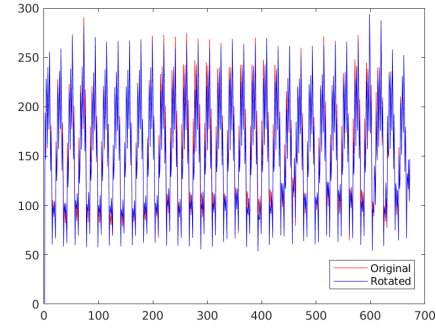
(a) Plot of CM descriptors



(b) Plot of LBP descriptors



(c) Plot of LOSIB descriptors



(d) Plot of SFTA descriptors

Figure 5.2: Descriptors extracted from one sample of the OASIS dataset. The descriptors are extracted again after the sample is rotated by 90° .

samples, the classification performance will suffer from over-fitting; hence, it will lead to low classification results on the testing samples.

Therefore, we can obtain the attribute representation of the averaged shearlets for the i^{th} slice as one vector $d_i = (d^1, d^2, \dots, d^j)$, where j represents the length of the vector. The final feature vector F for an MRI sample can be denoted as $F = (d_1, d_2, \dots, d_m)$, where m represents the number of slices of an MRI sample.

Texture-Based Descriptors in the 3D Shearlet Domain

Thereafter, a texture-based technique, i.e., CM, LBP, LOSIB, or SFTA, is applied to those averaged shearlet sub-bands that represent one MRI slice. The details of those methods are provided in Chapter 3. However, all descriptors of all slices for each method are then concatenated to form the feature vector of an MRI sample.

In the previous section, we described the process of reducing the shearlet sub-bands to only one averaged sub-band for each MRI slice. Obviously, a limitation of such an approach is that the directionality information is lost. To further investigate the rotation invariance of each texture-based method, one sample from the OASIS dataset is used to

form the feature vector for each method. Then, the same sample is rotated by 90° and used to obtain the descriptors in the averaged shearlet domain. In Figure 5.2, we observe that CM, LBP, and SFTA descriptors are rotationally invariant, but the LOSIB descriptors are not rotationally invariant. From Chapter 3, we know that the CM descriptors and LBP are rotationally invariant. Particularly, we obtain the mean of CM from four orientations as Kather et al. [103], which makes the descriptors invariant with respect to rotation.

Several fusion approaches have been proposed, and have shown varying levels of success in medical image classification [7, 103]. Since the underlying concept of each texture-based method is different with some methods capturing local but other global statistics, it is interesting to fuse feature vectors of different methods to observe any advantages in building a classification model. Therefore, we explore the following feature fusion:

- Fusion #1: (CM + LBP + LOSIB + SFTA) descriptors. This fusion is the same that was carried out in Chapter 3, but without phase coefficients.
- Fusion #2: (CM + SFTA) descriptors. This fusion was selected based on the high classification results obtained by these methods individually.

The setting of those texture-based techniques can be adjusted. Each texture-based method has a various number of hyper-parameters. Various parameters are tested for each texture-based technique to build a classification model. Then, based on the best performing model we select the hyper-parameter that maximizes the performance on the validation split of the dataset. Refer to Section 3.4 for more details. Therefore, it is not necessary that the parameters used for one MRI dataset are the same for another.

In this chapter, we utilize two different datasets (i.e., ADNI and OASIS). Prior to transforming the ADNI dataset, we pre-process each MRI sample using CAT12 [68]. The following pre-processing techniques are applied: skull-stripping, image normalization, and cropping (as shown in Figure 5.4). However, when the OASIS dataset is used, we only apply cropping. Then, each pre-processed MRI sample is transformed using the 3D shearlet. However, the results reported in this chapter for the two datasets are based on the following hyper-parameters:

- **ADNI:** for CM: we use a displacement of 1. For LBP: we use a radius (R) of 2 and a neighborhood (P) of 16. For LOSIB: we use a radius (R) of 5 and a neighborhood (P) of 16. For SFTA: we use a threshold of 1.
- **OASIS dataset:** for CM: we use a displacement of 2. For LBP: we use (R, P) = (3, 16). For LOSIB: we use a (R, P) = (3, 16). For SFTA: we use a threshold of 4.

5.2.2 Fine-tuning Pre-trained CNN Models for AD Classification

In this study, we examine pre-trained models on non-medical images, i.e., ImageNet [46]. Four pre-trained CNN architectures are fine-tuned for AD diagnosis as described in this section. Then, those pre-trained CNN models are leveraged for their capability as feature

extractors. After that, we explore the combination of deep features in an attempt to increase the classification performance while using SVM or DTB. It is worth noting that, unlike the shearlet-based approach, the MRI volumes are NOT pre-processed for skull-stripping, image normalization, or cropping. The CNN model rather is given full control to learn such pre-processing steps.

Fine-tuning a Pre-trained Model

Sharma and Mehra [192] demonstrated that fine-tuning a pre-trained model coupled with logistic regression as a classifier in the task of breast cancer histology image classification led to good classification performance. In this study, we propose to use four pre-trained CNN models that can handle the complexity of MRI datasets: SqueezeNet-v1.1 [93], MobileNet-v2 [186], Inception-v3 [209], and Xception [38]. Shallu et al. fine-tuned a pre-trained model using histology images that have the same number of channels as the input layer of the pre-trained model. In our case, the MRI sample consists of a different channel number as we stack the slices in depth. Consequently, each adopted CNN model will need to be adjusted in the following manner:

- The input layer needs to be changed to fit the dimensions of an MRI sample, i.e., $h \times w \times c$, where the number of channels of an MRI sample is $c > 3$. Since the input layer is pre-trained on RGB images, then the convolution layer which follows the input layer has a pre-trained weights matrix with a channel size = 3. Therefore, this convolution layer must be replaced to fit the channel size of the new input layer and trained from scratch.
- The specialized layers prior to the classification layer, typically convolution and pooling layers or dense layers, are replaced and trained from scratch using new filter sizes to fit the size of the preceding feature maps. Finally, given the dataset we are tackling is a binary problem, our classification layer which is also trained from scratch must have an output size of two.

It is worth noting that rest of pre-trained weights of a CNN model are kept and used as weight initializers. As such, they are fine-tuned as described in the following section.

CNN as Feature Extractor

A pre-trained deep model can be used to extract a feature vector per MRI slice, and then concatenating all feature vectors of all slices can form a single feature vector representing an MRI sample. Such a process will generate a very large number of descriptors per MRI sample. However, we incorporate necessary changes to a pre-trained DL model for MRI classification, as described in the previous section. We make use of state-of-the-art deep networks to learn the deep features of a stacked MRI sample simultaneously. A brief description of each model used in our study and the targeted feature maps as follows:

- **SqueezeNet-v1.1** consists of eight modules, namely Fire2 - Fire9. Each module has two expand layers, one of a kernel size of 1×1 and the second one of size 3×3 . Both expansions feed into a depth concatenation layer, generating feature maps of size $H \times W \times C$ for each module. In this model, we select module ‘Fire9’. We take the mean of each feature map generating a feature vector of size $1 \times 1 \times C$, which is finally flattened to form the feature vector of an MRI sample.
- **MobileNet-v2** has 16 blocks of two types: the first type consists of linear bottleneck operation, and a skip connection. Both blocks contain the following operations: convolution, batch normalization, and modified rectified linear unit (ReLU). The ReLU is modified to limit the maximum output with $\min(\max(x, 0), 6)$. In this model, we use the output of the average pooling (Avg Pool) layer, the deepest layer, just prior to the classification layer.
- **Inception-v3** is a model that typically consists of a module that maps input data into 3 or 4 different branches where each branch looks at the cross-channel correlations with a set of convolutions that has a size of 1×1 . Further processing these feature maps via another set of convolutions with filters of 3×3 or 5×5 . Then, all of these feature maps are concatenated. In this model, we use the output of the Avg Pool layer which is the deepest layer just before the classification layer.
- **Xception** is a model that is completely based on depthwise separable convolution layers. This model consists of 36 convolutional layers which are formulated into 14 blocks. Only the first and last blocks do not have residual connections around them, but the rest have a linear residual connection around them. Similar to the Inception-v3, we explore the Avg Pool layer in this model.

Thereafter, a standard classifier, namely SVM and DTB, is trained in the context of nested cross-validation (see Section 5.3.2).

Deep Features Fusion

Each CNN architecture is modeled differently to capture and process the input data through a sequence of operations that learns a function of the feature maps that are coming from previous layers. Furthermore, some models apply skip connections differently (e.g., Inception-v3 applies depth concatenation, but Xception applies element-wise addition) processing feature maps of the current module with the previous one. Other studies demonstrated success in combining deep features for medical image classification [174, 228]. Therefore, we explore the following feature fusion of various deep features that are obtained under different supervised DL models:

- DF Fusion #1: (SqueezeNet(Fire9) + MobileNet (Avg Pool) + Xception(Avg Pool) + Inception(Avg Pool)) descriptors.
- DF Fusion #2: MobileNet (Avg Pool) + Xception(Avg Pool) + Inception(Avg Pool) descriptors.

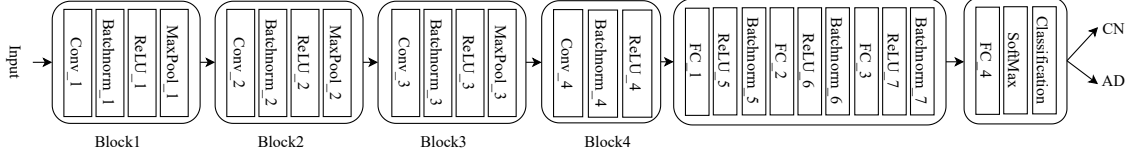


Figure 5.3: Our custom CNN model which is used to transform the combined shearlet-based descriptors with deep features.

5.2.3 Custom CNN

Although in the 3D shearlet transform we do not make full use of the multiple scales and directionalities, we still use averaged shearlets as a reduced form of the coefficients. This reduced averaged shearlet sub-band is used to obtain low-level image structures, e.g., utilizing SFTA, from which local and global statistics are computed. Alternatively, one can obtain high-level semantic deep features that are computed from different CNN models. Then, we investigate whether combining shearlet-based descriptors and deep features enhances the classification model, and fusing such attributes provides a more discriminative and effective classifier.

Therefore, our main idea explored in this chapter is to concatenate feature vectors from shearlet-based descriptors and deep features. A typical and simple approach of feature fusion is to concatenate both shearlet descriptors and deep features before feeding the new combined feature vector to a classifier. We will compare our method with this simple approach.

We propose to integrate the shearlet descriptors and deep features by transforming such features using a custom CNN model as shown in Figure 5.3. The building block of our custom CNN model consists of the following operations: convolution (Conv), batch normalization, Rectified Linear Unit (ReLU), and then max pooling. For the convolution layers, we use filter size of 7 for each convolution layer, but the number of filters is 8, 16, 32, and 64, respectively. The convolution layer is similar to a feature extraction step in the traditional pipeline, but the parameters of this layer can be learned with an optimizer. However, the filter size for the max-pooling layers is 1 with a stride of 2. A pooling layer is an important element of our CNN model for feature reduction and speeding-up the training process. Furthermore, to avoid over-fitting a dataset while training our model, we use batch normalization. Our building block is repeated four times and then followed by three fully connected layers that consist of 120 neurons and then one fully connected layer of 2 neurons corresponding to our binary classification problem. It is worth noting that our model is derived empirically, such that our custom CNN model can generalize on the validation data for the ADNI and OASIS datasets we use in our study.

We make use of our custom CNN model to fuse our hand-crafted descriptors and deep features. Previously, we examined the combination of features exclusively within the same domain, i.e., either hand-crafted descriptors together, or deep features together. In this section, we propose to combine both handcrafted and deep features altogether. The following fusions are compared:

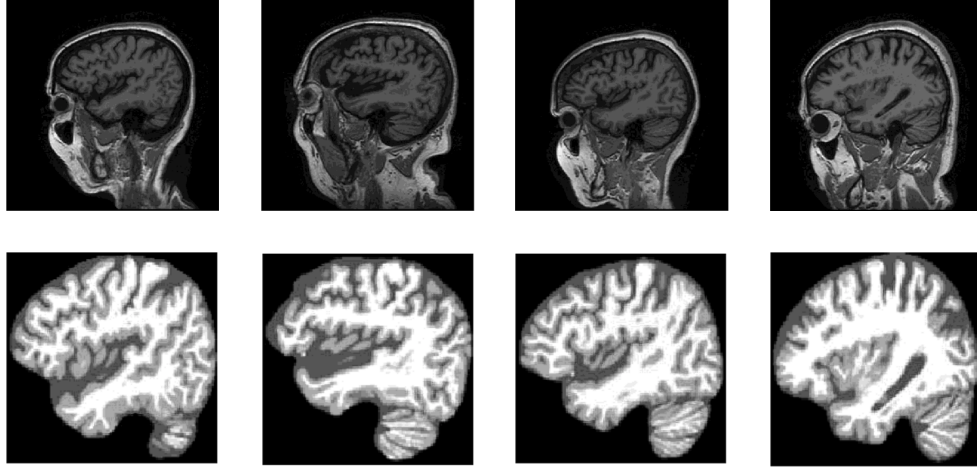


Figure 5.4: Examples of MRI slices before (top) and after (bottom) processing. Left to right: CN, sMCI, AD, and pMCI, respectively.

- All of our extracted hand-crafted descriptors are aggregated as one data matrix (i.e., X) then aggregated with our all deep features (i.e., Y). This fusion is general to all MRI brain scans.
- Alternatively, one selected set of hand-crafted descriptors with one selected set of deep features (i.e., each set is selected based on its performance while being used individually). This fusion is specific to an MRI brain scans dataset.

5.3 Experiments

5.3.1 Datasets

Alzheimer’s Disease Neuroimaging Initiative (ADNI) Data

We use the same data samples as Salvatore et al. [185]. The data is archived and can be accessed from Alzheimer’s Disease Neuroimaging Initiative (ADNI) data archive*. The dataset consists of 200 patients. The distribution of samples in this dataset is as follows: 50 patients with a stable diagnosis of cognitively normal (CN) during the 24 months follow-up. 50 patients with a stable diagnosis of mild cognitive impairment (sMCI), 50 patients with a stable diagnosis of Alzheimer’s disease (AD), and 50 patients with a MCI diagnosis that may progress to AD (pMCI). All patients had three serial MRI examinations at multiple points after the baseline: after 6, 12, and 24 months. However, since our target is to examine the capability of our technique to distinguish between Alzheimer’s patients from non-Alzheimer’s patients using MRI data, our focus in this study is on the baseline MRI scans of the patients.

*The ADNI data archive can be accessed via adni.loni.usc.edu.

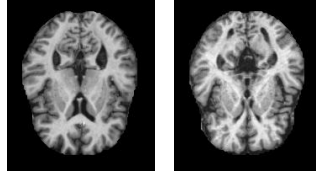


Figure 5.5: Examples of the OASIS MRI slices. CN on the left and AD on the right.

The structural MRI dataset is constructed at 1.5T using a T1-weighted sequence as maintained by the ADNI standard acquisition protocol [96]. All MRI images are pre-processed for a 3D-gradwarp geometry correction for gradient non-linearity, and a B1 intensity correction for non-uniformity. Examples of this dataset are shown in Figure 5.4.

We carry out the following classification task while using the ADNI dataset as Salvatore et al.: (CN + sMCI) vs. (pMCI + AD)[†].

Further, to evaluate the generalization of our proposed method, we retrieve a separate set of MRI data from the ADNI data repository. As such, 10 MRI volumes per category are retrieved (i.e., CN, sMCI, AD, and pMC), but these MRI volumes are of different time-points (i.e., not necessarily of baseline scans). Such a dataset of new subjects is not seen during the training process from start to end which will be used to test our proposed method at the very end.

To be able to use the ADNI data, we had to obtain approval for secondary use of this data from the Research Ethics Boards (REB) at the University of Ottawa. The Ethics file number is H-11-20-6270.

The Open Access Series of Imaging Studies (OASIS) Data

To further examine our proposed techniques, we retrieve another set of cross-sectional MRI data from the OASIS data repository[‡] for AD classification [137]. Examples of AD and CN slices are shown in Figure 5.5. In this experiment, we retrieve 100 samples of each CN and AD patients. We retrieve the same 200 subjects[§] that are used by Hon and Khan [85]. However, 10 examples out of each of the 100 samples from each category (i.e., CN and AD) are kept hidden from start to end. Then, these hidden examples are used at the very end to examine the generalization of our proposed method. Hence, we use only 90 examples of each category to train and validate our proposed techniques.

The OASIS data repository provides MRI scans as individual slices for a patient. However, we instead deal with the data as a volume for each patient’s slices by converting them into the neuroimaging informatics technology initiative (NIFTI) format. Again, we apply a binary classification while using this data: CN vs. AD.

[†]More information about the dataset and patients IDs can be accessed at (<https://github.com/christiansalvatore/Salvatore-200Longitudinal>).

[‡]The OASIS data repository can be accessed via <http://www.oasisbrains.org>.

[§]More information about the patients can be found at: https://github.com/marciahon29/Ryerson_MRP.

5.3.2 Nested Cross-validation (CV)

In this study, we utilize 5-fold nested cross-validation for hyper-parameter optimization. While using a 5-fold nested CV, the data is split into 5 subsets of equal sizes. Therefore, 4 out of 5 subsets are used for training and validation for hyper-parameter optimization in an inner loop. Then, the last subset is used for testing the performance of a classifier in an outer loop.

For each optimization round, the set of hyper-parameters are tuned to maximize the geometric mean of the classifier in the inner loop. For each optimization round, the classification performance is estimated in the outer loop in terms of accuracy, sensitivity, specificity, and geometric mean.

The number of samples that are used in our nested CV is as follows:

- Given that the number of MRI volumes from the ADNI data is 200, i.e., 100 CN + sMCI, and 100 pMCI + AD, for each round of the nested CV, the number of MRI samples that are used to train a classifier is 128; the number of samples that are used in the inner loop for optimization is 32; the number of MRI samples that are used to test the performance in the outer loop is 40.
- Given that the number of MRI volumes from the OASIS data is 180, i.e., 90 CN and 90 AD, for each round of the nested CV, the number of MRI samples that are used to train a classifier is 115; the number of samples that are used in the inner loop for optimization is 29 but, except for one round, a classifier is trained with 116 samples and validated with 28 samples. However, the number of samples that are used to test the performance in the outer loop is 36.

We use 5-fold nested cross-validation, i.e., the above procedure is repeated 5-times, and then across the 5 rounds, the average of common classification performance in the outer loop is reported. We use the following metrics:

- Accuracy (ACC) is the ratio of correctly classified over total number of samples, or $ACC = (TP + TN) / (TP + TN + FP + FN)$
- Sensitivity (Sen) is the ratio of true positives over actual positives, or $Sen = TP / (TP + FN)$
- Specificity (SP) is the ratio of true negatives over actual negatives, or $SP = TN / (TN + FP)$
- Geometric Mean $GM = \sqrt{Sen \times SP}$

In the following Section 5.4, we first discuss results classifying the Alzheimer’s disease while using shearlet-based descriptors individually and jointly, and report results while using two classical classifiers (i.e., SVM and DTB). Similarly, we report the performance of the SVM and DTB classifier when being trained based on the various deep features

Table 5.1: ADNI classification performance using shearlet-based descriptors for (CN + sMCI) vs. (pMCI + AD).

Best performance is highlighted.

Method		ACC	Sen	SP	GM
(a): Shearlet-based methods for AD Classification					
CM	SVM	0.6800±0.0542	0.6700±0.1037	0.6900±0.0418	0.6778±0.0570
	DTB	0.6850±0.0335	0.7100±0.0742	0.6600±0.0418	0.6831±0.0332
LBP	SVM	0.6700±0.0570	0.7000±0.0791	0.6400±0.0822	0.6674±0.0571
	DTB	0.6300±0.0671	0.6300±0.0758	0.6300±0.0908	0.6284±0.0678
LOSIB	SVM	0.6450±0.0326	0.6700±0.0447	0.6200±0.0570	0.6436±0.0324
	DTB	0.6400±0.0379	0.6400±0.0652	0.6400±0.0652	0.6382±0.0389
SFTA	SVM	0.7050 ± 0.0447	0.7300 ± 0.1151	0.6800±0.0570	0.7009±0.0458
	DTB	0.6750±0.0433	0.6300±0.0837	0.7200 ± 0.0447	0.6718±0.0460
(b): Combining shearlet-based techniques					
Fusion #1	SVM	0.6750±0.0500	0.6800±0.0671	0.6700±0.0837	0.6730±0.0510
	DTB	0.7050 ± 0.0542	0.7200±0.0671	0.6900±0.0652	0.7040 ± 0.0541
Fusion #2	SVM	0.6750±0.0354	0.6700±0.0758	0.6800±0.0274	0.6738±0.0356
	DTB	0.7050±0.0597	0.7100±0.1025	0.7000±0.0500	0.7033±0.0599

computed from different fine-tuned CNN architectures. Our classical ML classifiers are trained and validated in the context of nested cross-validation.

However, given the fact that training a deep learning model is time-consuming, our custom CNN models are trained from scratch while using 80% for training and 20% for validation. Then, the best performing model based on the validation split is tested on the unseen data that is hidden from start to end while fusing shearlet-based descriptors with deep features.

5.4 Results and Analysis

5.4.1 Classification Results Using Shearlet-Based Descriptors

First, we report results while using shearlet-based techniques for feature extraction (see Section 5.2.1). The performance of two classifiers, i.e., SVM and DTB, are reported in Tables 5.1, and 5.2 for the ADNI and OASIS datasets that are used in our study. In our study, 5-fold nested cross-validation is used to obtain the classification results of the aforementioned classifiers.

When using the ADNI dataset, our proposed descriptors extracted from the 3D shearlet

Table 5.2: OASIS classification performance using shearlet-based descriptors for CN vs. AD.

Best performance is highlighted.

Method		ACC	Sen	SP	GM
(a): Shearlet-based methods for AD classification					
CM	SVM	0.6889 $\pm$ 0.1028	0.6444 $\pm$ 0.1394	0.7333 $\pm$ 0.0724	0.6860 $\pm$ 0.1066
	DTB	0.6889 $\pm$ 0.0602	0.6333 $\pm$ 0.1152	0.7444 $\pm$ 0.0843	0.6827 $\pm$ 0.0613
LBP	SVM	0.6889 $\pm$ 0.0819	0.6556 $\pm$ 0.1542	0.7222 $\pm$ 0.0786	0.6823 $\pm$ 0.0917
	DTB	0.6667 $\pm$ 0.0810	0.6333 $\pm$ 0.1152	0.7000 $\pm$ 0.0930	0.6628 $\pm$ 0.0847
LOSIB	SVM	0.6556 $\pm$ 0.0421	0.5889 $\pm$ 0.0843	0.7222 $\pm$ 0.1039	0.6479 $\pm$ 0.0380
	DTB	0.6278 $\pm$ 0.0824	0.6444 $\pm$ 0.1152	0.6111 $\pm$ 0.1179	0.6229 $\pm$ 0.0850
SFTA	SVM	0.6222 $\pm$ 0.0505	0.6444 $\pm$ 0.1009	0.6000 $\pm$ 0.0724	0.6185 $\pm$ 0.0506
	DTB	0.6444 $\pm$ 0.0819	0.6667 $\pm$ 0.1303	0.6222 $\pm$ 0.0609	0.6418 $\pm$ 0.0815
(b): Combining shearlet-based techniques					
Fusion #1	SVM	0.6722 $\pm$ 0.0692	0.6111 $\pm$ 0.1620	0.7333 $\pm$ 0.0824	0.6614 $\pm$ 0.0798
	DTB	0.7000 $\pm$ 0.0534	0.6222 $\pm$ 0.1204	0.7778 $\pm$ 0.0556	0.6916 $\pm$ 0.0625
Fusion #2	SVM	0.6556 $\pm$ 0.0697	0.5444 $\pm$ 0.1542	0.7667 $\pm$ 0.0609	0.6392 $\pm$ 0.0843
	DTB	0.7000 $\pm$ 0.0819	0.6111 $\pm$ 0.1620	0.7889 $\pm$ 0.0609	0.6888 $\pm$ 0.0895

coefficients can achieve reasonable accuracy between 68% to 70% when fed to an SVM or DTB classification model as shown in Section (a) of Table 5.1. We notice that the highest accuracy is 0.7050 obtained while classifying this dataset utilizing SFTA descriptors coupled with an SVM classifier with a corresponding sensitivity = 0.7300, specificity = 0.6800, and geometric mean = 0.7009. On the other hand, when we use Fusion #2 that combines CM and SFTA descriptors, the fusion provides a balance between sensitivity = 0.7100 vs. specificity = 0.7000 as presented in Section (b) of Table 5.1 while using a DTB model. Accuracy remains the same as utilizing SFTA descriptors alone.

Similarly, we apply our 3D shearlet technique on the OASIS dataset. While using the OASIS dataset, our technique achieves accuracies between 62% to 68% when fed to a classification model as shown in Section (a) of Table 5.2. The highest obtained accuracy is 0.6889 while classifying the dataset utilizing co-occurrence matrix textural features coupled with an SVM classifier with a corresponding sensitivity of 0.6444, a specificity of 0.7333, and a geometric mean of 0.6860. On the other hand, when we combine CM, LBP, LOSIB, and SFTA in descriptors Fusion #1, the fusion boosts the accuracy to 0.7000 while using a DTB model as presented in Section (b) of Table 5.2.

As a conclusion of this section, we compute the t-test statistical significance of best performing models when descriptors are used individually vs. jointly based on geometric mean values. We can observe from Table 5.1 for the ADNI data that SFTA alone yields

Table 5.3: ADNI classification performance using deep features for (CN + sMCI) vs. (pMCI + AD).

Best performance is highlighted.

Method		ACC	Sen	SP	GM
SqueezeNet	SVM	0.8400±0.0859	0.8569±0.0878	0.8218±0.1102	0.8378±0.0831
	DTB	0.8950±0.0570	0.8774±0.0602	0.9132±0.0909	0.8939±0.0543
MobileNet	SVM	0.9050±0.0411	0.8948±0.0898	0.9063±0.0628	0.8987±0.0443
	DTB	0.9100 ± 0.0379	0.9161 ± 0.0649	0.8964±0.0563	0.9052 ± 0.0381
Xception	SVM	0.9000±0.0586	0.8744±0.1016	0.9217±0.0645	0.8963±0.0659
	DTB	0.9050±0.0737	0.8839±0.1124	0.9241 ± 0.0846	0.9021±0.0789
Inception	SVM	0.8650±0.0652	0.8861±0.0779	0.8410 ± 0.0572	0.8631±0.0656
	DTB	0.8800±0.0818	0.8839±0.1018	0.8727±0.0822	0.8776±0.0848
DF Fusion #1	SVM	0.8700±0.0447	0.8756±0.0707	0.8608±0.0818	0.8665±0.0483
	DTB	0.8800±0.0818	0.8839±0.1018	0.8727±0.0822	0.8776±0.0848
DF Fusion #2	SVM	0.8650±0.0487	0.8661±0.0603	0.8608±0.0818	0.8622±0.0522
	DTB	0.8800±0.0818	0.8839±0.1018	0.8727±0.0822	0.8776±0.0848

a GM value of 0.7009 but when we use Fusion #1 of shearlet descriptors a slightly higher GM value of 0.7040 is obtained. Nevertheless, when a t-test statistical significance of those two models is computed, we obtain a t-value of -0.0972 and a p-value of 0.4625 which we conclude that the result is not significant at $p < 0.05$. Similarly for the OASIS dataset in Table 5.2, CM features alone yields the highest GM value of 0.6860 but Fusion #1 leads to a GM value of 0.6916. Again, we conclude the result of classifying the OASIS dataset while using CM vs. combined features is not significant at $p < 0.05$ where the t-value is -0.1024 and the p-value is 0.4605.

5.4.2 Classification Results using Fine-tuned Pre-trained CNNs

In this section, we use the ADNI and OASIS datasets to fine-tune a CNN model. In order to be able to fine-tune a CNN model based on an MRI volume, we need to adjust the model as detailed in Section 5.2.2. To fine-tune a CNN model, the dataset used in this study is divided into 80% for training and 20% for validation. The purpose of such a division is to select and monitor the hyper-parameters and the model performance, respectively. We choose a momentum value of 0.9 and a small learning rate of 0.0001 while utilizing stochastic gradient descent for model fine-tuning [145]. Furthermore, fine-tuned pre-trained CNN models are trained for a maximum of 60 epochs with mini-batches of size 10. In this study, we re-train various CNN models: SqueezeNet, MobileNet, Xception, and Inception. After fine-tuning a model, each model is then used to generate a feature vector for each MRI volume. Those extracted feature vectors of each CNN model are used to train two classifiers, i.e., SVM and DTB, in the context of 5-fold nested cross-validation as described

Table 5.4: OASIS classification performance using deep features for CN vs. AD.

Best performance is highlighted.

Method		ACC	Sen	SP	GM
SqueezeNet	SVM	0.8944±0.0497	0.8603±0.0492	0.9313±0.0714	0.8942±0.0422
	DTB	0.8833±0.0534	0.8637±0.0899	0.9146±0.0885	0.8861±0.0480
MobileNet	SVM	0.9167±0.0196	0.9034±0.0688	0.9367±0.0579	0.9184±0.0225
	DTB	0.9278 ± 0.0248	0.9146±0.0703	0.9467 ± 0.0542	0.9291 ± 0.0266
Xception	SVM	0.9000±0.0541	0.9034±0.0688	0.9035±0.1113	0.9010±0.0562
	DTB	0.8889±0.0481	0.8909±0.0500	0.8929±0.1169	0.8894±0.0513
Inception	SVM	0.8889±0.0589	0.9091±0.0539	0.8723±0.1071	0.8889±0.0622
	DTB	0.9056±0.0639	0.8966±0.0529	0.9129±0.1266	0.9023±0.0653
DF Fusion #1	SVM	0.8889±0.0520	0.9034±0.0688	0.8823±0.1068	0.8906±0.0568
	DTB	0.9000±0.0609	0.8966±0.0529	0.9029±0.1198	0.8975±0.0631
DF Fusion #2	SVM	0.9000±0.0639	0.9216 ± 0.0676	0.8823±0.1068	0.9001±0.0679
	DTB	0.9000±0.0609	0.8966±0.0529	0.9029±0.1198	0.8975±0.0631

in Section 5.3.

Tables 5.3 and 5.4 show results of various CNN models when used as feature extractors after being fine-tuned while using the ADNI or the OASIS dataset, respectively. Fine-tuning a pre-trained MobileNet leads to high classification results while using a DTB model when trained in the five-fold nested CV fashion for both datasets that were used in our study.

We also notice that when we explore two variations of deep features fusion to classify both datasets (i.e., the ADNI and OASIS), the classification results did not improve. Further, we compute the t-test statistical significance of the results of MobileNet vs. deep features fusion #2 for classifying the ADNI dataset, the result is not significant at $p < 0.05$ where the obtained t-value and p-value are 0.6631 and 0.2629, respectively. Similarly, when the t-test is computed for the performance of both MobileNet vs. deep features fusion #2 for classifying the OASIS dataset, the result is not significant at $p < 0.05$ with a t-value of 0.8885 and p-value of 0.2001.

5.4.3 Classification Results While Combining Shearlet-based Descriptors with Deep Features

In the two previous sections, we provided a comprehensive examination of our shearlet-based descriptors and deep features. In Table 5.5 and Table 5.7, we explore the combination of shearlet-based features with deep features when being fed to an SVM or DTB classifier. When we use all of our hand-crafted descriptors with deep features in Fusion #1 to classify the ADNI dataset, we obtain an accuracy of 0.9050 with an SVM classifier. However, a higher classification accuracy of 0.9350 is achieved with an SVM model while selecting only

Table 5.5: ADNI classification performance while combining shearlet-based descriptors with deep features for (CN + sMCI) vs. (pMCI + AD).

Best performance is highlighted. SVM and DTB classifiers are trained in the context of 5-fold nested CV, but CNN is trained while using 80% - 20% for training and validation, respectively.

Method		ACC	Sen	SP	GM
(a): Shearlet-based descriptors with deep features without using the custom CNN					
Fusion #1	SVM	0.9050±0.0274	0.9100±0.0418	0.9000±0.0500	0.9044±0.0276
	DTB	0.8750±0.0306	0.8900±0.0418	0.8600±0.0548	0.8742±0.0311
Fusion #2	SVM	0.9350 ± 0.0285	0.9400 ± 0.0418	0.9300±0.0447	0.9345 ± 0.0283
	DTB	0.8950±0.0597	0.9100±0.0224	0.8800±0.1151	0.8933±0.0618
(b): Shearlet-based descriptors with deep features while using the custom CNN					
Fusion #1	CNN	0.9000	0.8500	0.9500	0.8986
Fusion #2	CNN	0.9000	0.9000	0.9000	0.9000

Table 5.6: Using unseen data samples of ADNI to test the generalization of our proposed techniques.

Best models are determined based on the highest geometric mean that is computed in Table 5.5.

Method		ACC	Sen	SP	GM
(a): Shearlet-based descriptors with deep features without using the custom CNN					
Fusion #2	SVM	0.6450±0.0371	0.4800±0.0671	0.8100±0.0418	0.6221±0.0425
(b): Shearlet-based descriptors with deep features while using the custom CNN					
Fusion #2	CNN	0.7000	0.6000	0.8000	0.6928

one of our shearlet-based descriptors (i.e., SFTA) and one of our deep features based on MobileNet. The selection is made based on the high performance when used individually. Furthermore, we examine the generalization of this high performing model on an unseen dataset that was hidden from start to end (as shown in Table 5.6). As such, we obtain an accuracy of 64.50% with a corresponding sensitivity of 0.4800, a specificity of 0.8100, and a geometric mean of 0.6221.

Similarly, we apply our proposed combination to classify the OASIS dataset. We obtain the highest classification accuracy of 90.00% while using a DTB when we combine all of our shearlet-based descriptors with all of our deep features. However, when we only select one of our shearlet-based descriptors and one of our deep features (i.e., the ones that achieved the highest accuracy on their own), the obtained classification accuracy is 92.78%; therefore, this model was selected to be tested on unseen data (i.e., hidden from the beginning to

Table 5.7: OASIS classification performance while combining shearlet-based descriptors with deep features for CN vs. AD.

Best performance is highlighted. SVM and DTB classifiers are trained in the context of 5-fold nested CV, but CNN is trained while using 80% - 20% for training and validation, respectively.

Method		ACC	Sen	SP	GM
(a): Shearlet-based descriptors with deep features without using the custom CNN					
Fusion #1	SVM	0.8944±0.0412	0.9000±0.0724	0.8889±0.0878	0.8922±0.0422
	DTB	0.9000±0.0421	0.9000±0.0724	0.9000±0.0824	0.8981±0.0424
Fusion #2	SVM	0.9278±0.0541	0.9111±0.0497	0.9444 ± 0.0680	0.9274±0.0539
	DTB	0.8944±0.0692	0.8889±0.0680	0.9000±0.0724	0.8944±0.0691
(b): Shearlet-based descriptors with deep features while using the custom CNN					
Fusion #1	CNN	0.8611	0.9130	0.7692	0.8381
Fusion #2	CNN	0.9444	0.9565	0.9231	0.9397

Table 5.8: Using unseen data samples of OASIS to test the generalization of our proposed techniques.

Best models are determined based on the highest geometric mean that is computed in Table 5.7.

Method		ACC	Sen	SP	GM
(a): Shearlet-based descriptors with deep features without using the custom CNN					
Fusion #2	SVM	0.8000 ± 0	0.7000 ± 0	0.9000 ± 0	0.7937 ± 0
(b): Shearlet-based descriptors with deep features while using the custom CNN					
Fusion #2	CNN	0.8000	0.7000	0.9000	0.7937

end), the accuracy on the unseen data is 80.00% with a corresponding sensitivity of 0.7000, a specificity of 0.9000, and a geometric mean of 0.7937 (as shown in Table 5.8).

In an attempt to reach a better classification model, we have built a custom CNN model (see Section 5.2.3). Such a model takes a feature vector as an input which will be further processed end-to-end to obtain a prediction. Such a custom CNN model is trained from scratch while utilizing 80% of the dataset for training and 20% for validation. The purpose of such a division is to select and monitor the hyper-parameters and the model performance, respectively. As a result, we choose a regularization value of 0.005 and a small learning rate of 0.001 while utilizing adaptive moment estimation (Adam) [106]. In addition, our custom CNN models are trained for 200 epochs with mini-batches of size 10.

Similar to before, we examine our custom CNN model while concatenating all of our shearlet-based descriptors with our deep features in Fusion #1. Alternatively, we examine

our custom CNN model while using only one of each domain, i.e., one from the shearlet-based descriptors and one from the deep features, that achieved the highest results individually. In Table 5.5, we observe that the validation geometric mean for classifying the ADNI dataset with the custom CNN for Fusion #2 is 0.9000. Therefore, we select Fusion #2 as the best model and test this model on the unseen dataset which leads to an accuracy of 70.00% with a corresponding sensitivity of 0.6000, a specificity of 0.8000, and a geometric mean of 0.6928 as shown in Table 5.6. Similarly in Table 5.7, when we use only co-occurrence textural features that are computed in the shearlet domain combined with MobileNet deep features in Fusion #2 to train and validate our custom CNN model on the OASIS scans, we obtain a validation geometric mean of 0.9397. This model is then tested on the unseen data to examine its generalization. The achieved accuracy on the unseen data is 80.00% with a corresponding sensitivity of 0.7000, a specificity of 0.9000, and a geometric mean of 0.7937 as shown in Table 5.8.

It is worth noting that our custom CNN provides higher classification results on the unseen dataset than using SVM or DTB. This is an indication that it is less susceptible to over-fitting. There is a gap between the validation results and the test results, especially on the ADNI datasets. As such, the validation accuracy and testing accuracy while using the ADNI dataset are 90.00% and 70.00%, respectively. However, as it has been indicated in the dataset section, the ADNI MRI volumes that are used for training/validation are of the baseline visits of the patients, but our unseen data subjects are of patients of different time-points. Thus, one can view our approach as effective to classify AD brain scans of patients that are not necessarily of the same time-point.

5.4.4 State-of-the-art Results

Many studies have been developed for the AD classification using traditional ML and DL. In Table 5.9, we provide a summary of several state-of-the-art classification performances of methods using the ADNI and OASIS datasets. Although we apply the classification task of (CN + sMCI) vs. (pMCI + AD) while using the ADNI dataset, it can be compared with the classification task of (sMCI vs. pMCI) [185].

Particularly, we compare the performance of our method on the ADNI dataset with results that were reported by Salvatore et al. [185]. We conduct our experiments for training a classifier while using the same patients' data samples of the ADNI dataset as Salvatore et al. When using the baseline of the ADNI dataset, Salvatore et al. reported the highest accuracy of 85%, with a corresponding sensitivity of 0.83, and a specificity of 0.87. This performance was achieved while using (PCA + FDR) extracted from the MRI samples with neuropsychological data. However, when only PCA features were used, the classification accuracy, sensitivity, and specificity were 0.72, 0.69, and 0.75, respectively. The classification performance improved with the neuropsychological data, but the authors did not report the performance while using neuropsychological data only. In comparison to Salvatore et al., our highest achieved accuracy is 93.50% while combining shearlet-based descriptors and deep features that are extracted from the MRI samples, coupled with an SVM model. When we examine our combined descriptors with a custom CNN model,

examined in a split of 80% for training and 20% for validation, it led to an accuracy of 90.00%. Additionally, we do not use neuropsychological data. It is also worth noting that, we examine the generalization of our best performing custom CNN model, i.e., achieved the highest geometric mean, on unseen data. Such a model led to an accuracy of 70.00%.

Furthermore, we also report the performance of a study done by Hon and Khan [85]. Our experiments are conducted in a similar manner for training a classifier while using the same patients' data samples of the OASIS dataset. Hon and Khan fine-tuned two CNN models while using each slice of an MRI as an independent example. When VGG16 was fine-tuned, they achieved an accuracy of 92.3%. However, a better classification accuracy of 96.25% was achieved while using Inception V4.

Similarly, Yagis et al. [233] explored fine-tuning DL models using the same data samples that were used by Hon and Khan. They fine-tuned VGG-16 and ResNet-50 for AD classification, where only the last three layers were fine-tuned, and other layers were frozen. Each slice of an MRI was considered as an independent labeled example from the OASIS dataset. Also, some slices were discarded based on image entropy. When the split of slices was done randomly where the same patient's slices can appear either in the training or the testing dataset, the highest reported classification accuracy was 92.5% while using ResNet-50. When ResNet-50 was trained in a subject-based split where slices of a patient could appear only in the training or testing split exclusively, the highest classification accuracy was 67.1%.

Both Yagis et al. [233], and Hon and Khan [85] trained and tested their models in 5-fold CV, but we test our model using 5-fold nested CV. Another difference in our proposed method is that we use each MRI sample of the OASIS data as a volume. However, our highest classification accuracy is 92.78% while combining shearlet-based descriptors and deep features, coupled with an SVM model. Also, we examine our combined descriptors with a custom CNN model, examined in a split of 80% for training and 20% for validation, which achieved an accuracy of 94.44%. Furthermore, our best performing custom CNN model trained on the OASIS dataset, i.e., achieved the highest geometric mean on the validation set, is examined for its generalization on unseen data. Such a model led to an accuracy of 80.00%.

In our comparison Table 5.9, only the study by Pan et al. [162] used a separate set for testing a model generalization among the studies that we have reviewed.

Another observation from the comparison table: when the OASIS dataset was used, some authors reported the performance based on the random split of slices. This is a drastic issue since results can be over-optimistic because of possible 'information leakage'. As has been indicated earlier, Yagis et al. [233] demonstrated that a subject-based split is more appropriate, but most likely, a classifier can be expected to perform more badly. It is worth pointing to studies conducted by Mendoza et al. [141] and Wang et al. [223]. Both of these studies selected a single slice to train and validate their respective models. Arguably, one slice might not be sufficient for predicting AD since the practice in the hospitals by a radiologist is using multiple slices [98]. In addition, Wang et al. mentioned in their study that their 8-layer CNN model was empirically driven, but the CNN model was adjusted to maximize the classification performance on the validation split. Therefore,

the chosen model and hyper-parameters are then biased to maximize the performance on the validation set.

Table 5.9: Comparative analysis of related work for AD classification.

Authors	Training Protocol	Separate Set for Testing	Subject	*RD	ACC	Sen	SP
Reported performance of methods on the ADNI dataset for (sMCI vs. pMCI) classification							
Nanni et al. [147]	20-fold CV.	×	134 sMCI + 76 pMCI	×	0.671	0.355	0.865
Salvatore et al. [184]	20-fold nested CV.	×	134 sMCI + 76 pMCI	×	0.66	-	-
Feng et al. [61]	10-fold CV.	×	160 sMCI + 120 pMCI	×	0.7448	0.7111	0.7664
Lian et al. [125]	2-fold CV.	×	465 sMCI + 205 pMCI	×	0.81	0.53	0.85
Pan et al. [162]	5-fold CV.	✓	134 sMCI + 76 pMCI	×	0.62	-	-
Ours	5-fold nested CV.	✓	50 CN + 50 sMCI + 50 pMCI + 50 AD	×	0.9350	0.9400	0.9300
Reported performance of methods on the OASIS dataset for CN vs. AD							
Jha et al. [98]	10-fold CV.	×	98 CN + 28 ADs	×	0.9006	0.9200	0.8778
Wang et al. [223]	49 subjects of each class were used for training, and the other 49 subjects from each class were used for testing.	×	98 CN + 28 AD + 70 AD (privately collected).	×	0.9765	0.9796	0.9735
Yagis et al. [233]	5-fold CV.	×	100 CN + 100 AD	✓	0.925	-	-

Continues on the next page

Table 5.9: Comparative analysis of related work for AD classification.

Authors	Training Protocol	Separate Set for Testing	Subject	*RD	ACC	Sen	SP
Hon and Khan [85]	5-fold CV.	×	100 CN + 100 AD	✓	0.9625	-	-
Mendoza et al. [141]	134 samples were used for training and 40 samples were used for testing.	×	87 CN + 87 AD	×	0.90	0.95	0.85
Ours	80% of the data used for training and 20% used for validation.	✓	90 CN + 90 AD	×	0.9444	0.9565	0.9231

* Random division (RD) of slices that potentially the same patient slices can appear in both training and testing datasets.

5.5 Summary

MRI datasets present a major challenge when used for the automatic classification of Alzheimer’s disease: having to deal with high-dimensional data. This creates the problem of dealing with a large number of descriptors per MRI sample, with some representing irrelevant attributes. Furthermore, with a low number of training samples, it can lead to over-fitting and to a less effective classification model. Therefore, we rigorously evaluate our proposed techniques while utilizing a commonly used training and validation protocol. Most importantly, we further verify the capability of our technique to generalize on new emerging data samples, i.e., we utilize a separate testing set of MRI volumes.

We have explored our previously successful descriptors on histopathological tissue types. We have applied necessary adjustments to achieve classification performance to be on par with state-of-the-art results. Instead of transforming each slice using 2D shearlets, we use 3D Shearlets which take into account adjacent slices. Then, we showed that using only the averaged shearlet coefficients from all directions can lead to an effective classification model for AD disease classification.

When it comes to deep features, it is obvious that fine-tuning a pre-trained model can lead to superior performance. In this chapter, we have conducted a comprehensive analysis using deep learning strategies. Again, we had to adjust the pipeline of our previous approach (i.e., Chapter 4) to work with MRI scans. As such, we have leveraged a CNN model to compute very compact attributes of high-dimensional MRI samples. First, we

fine-tune a pre-trained deep learning model. The classification layer of a CNN model is removed and instead, SVM and DTB classifiers are used. We learn beneficial knowledge from stacked MRI slices. Each CNN model learns a very compact number of descriptors of each MRI sample, which is effective in accomplishing superior classification results.

Finally, in an attempt to further improve the classification results, a custom CNN model is used to fuse and transform the feature representation while using both shearlet-based descriptors and deep features.

Chapter 6

Conclusion

The main goal of this thesis is to provide an effective and efficient framework to classify medical images. To pursue this goal, we have proposed two pipelines, either using hand-crafted techniques based on shearlet descriptors or deep learning strategies. Hand-engineered descriptors involve the development of an algorithm that is devised to carefully extract useful knowledge and attributes from the content of medical images. On the other hand, deep learning models are trained directly based on a large number of raw input images, generally, using non-medical images. We use such models as our starting point.

These DL models are examined on four public histopathological datasets of different sizes. For instance, Warwick-QU contains a total of 165 samples, yet our DL models led to remarkable classification results. In Figure 4.2, we show the performance of two DL models, namely, SqueezeNet and ResNet, while varying the number of tissues per class. Both of these models have an accuracy of approximately 90% while using only 200 tissue images per class that are available from the multi-class Kather’s dataset.

Both of our techniques are examined on images of histopathological tissues. In our traditional learning-based techniques, our modeling is based on shearlet-based descriptors, which provides better results than many existing techniques on four standard histopathological datasets. Therefore, it is expected to generalize to other similar texture-based datasets. Furthermore, we have provided state-of-the-art results while using deep learning strategies. Most importantly, we have demonstrated that fine-tuning deep learning models do not require aggressive data augmentation.

To further explore our proposed techniques, we have examined them on MRI in radiology. In addition to the challenges presented with histopathological images, the MRI contains a large number of pixels as each sample contains many 2D image slices. As a result, we needed to adjust our techniques to work successfully with MRI sample classification. Therefore, our shearlet-based feature representation is computed using the 3D shearlet transform. On the other hand, we do not use pre-trained deep learning models as unsupervised feature extractors because we would have to feed each slice independently to obtain a feature vector per slice. Therefore, we fine-tune a pre-trained model using the MRI samples and then use it as a feature extractor for the whole MRI sample. We obtain a very compact number of descriptors which we use to train a classical ML classifier.

Furthermore, we combine our shearlet-based descriptors with deep features for Alzheimer’s disease classification.

As recommendations for practitioners for building medical image classification system:

- Pre-trained CNN models as unsupervised feature extractors is a desirable approach when dealing with 2D medical images, e.g., histopathological images. When it comes to data like MRI sample, which contains a large number of slices, it is rather more practical to fine-tune a DL model and used it as a feature extractor to obtain a compact feature vector per MRI sample. Then, these deep features can be used to train a classical classifier.
- Given that deep learning models might require demanding computational resources, texture-based methods, particularly CM textural features and LBP, with spectral methods can also deliver remarkable results. Generally speaking, this can work well with 2D datasets. When it comes to 3D datasets (e.g., MRI data), over-fitting might be a problem. A large number of descriptors will be computed. With a limited number of samples, this can potentially lead to over-fitting. This can be avoided by utilizing feature reduction techniques.

6.1 Summary

In this Ph.D. thesis, each of our proposed techniques for medical image classification consists of several components. Our hand-crafted techniques start from feature extraction, feature selection, to classification. When we use deep learning techniques, the learning process is end-to-end including the classification task. However, when a deep learning model is used as a feature extractor, then we need to apply feature extraction, then feature selection, and classification. Both of our techniques are used to examine medical images in the field of pathology and radiology.

In **Chapter 1**, we first provide an overview of two major fields in medicine, namely pathology and radiology. Both fields in medicine are receiving a great amount of interest given the fact that both are essential in identifying potentially lethal diseases. Radiology is inherently in digital form, while histological slides first need to be digitized.

In **Chapter 2**, we have provided an extensive summarization of available histopathological datasets. Further, we have surveyed studies that have used these datasets and reported their classification performance. We categorize our review of such techniques into traditional and deep learning-based methods. We have also identified the common procedure of training and testing a CAD. We further have a review of CAD techniques that can be applied to image classification in radiology. It is worth noting that each modality presents itself with different attributes of complexity, yet CAD techniques can be extended to other imaging modalities with necessary changes. Being able to identify the challenges presented for CAD systems, we have proposed novel techniques that are capable of delivering robust classification performance. For instance, histopathological images are challenging because

of variations due to different applications of staining components. Therefore, many authors proposed to apply color deconvolution to highlight the cellular level. We instead have used the complex shearlet coefficients to compute phase and magnitude coefficients which in return generated robust descriptors (see Chapter 3). We showed that while using deep learning, there is no need to train from scratch or to use aggressive data augmentation to achieve the state of the art results (see Chapter 4). Additionally, we showed that our proposed hand-crafted and deep learning techniques can be extended to MRI sample classification with necessary modification (see Chapter 5).

In **Chapter 3**, instead of computing the textural features in the spatial domain, we exploit the multi-directional and multi-resolution complex shearlet transform to compute our descriptors. With complex coefficients, we study the effectiveness of the shearlet coefficients when using the magnitude vs. relative phase (RP) components. Generally, we have found that the combination of magnitude and RP coefficients is beneficial for robustness and helps in achieving competitive results on public histopathological image datasets. Furthermore, our computed feature sets are based on methods that capture local and global statistics: co-occurrence textural features, local binary patterns, local oriented statistic information booster, and segmentation-based textural features calculated in the complex shearlet domain. After that, we use PCA to obtain a reduced set of our proposed combined feature space in the shearlet domain. On some datasets, this reduced feature set can maintain or increase classification performance with traditional machine learning. We also show that the machine learning method has only limited influence on the results and hence it is possible to use a DTB if desired because decisions of this classifier are considered interpretable. Our proposed method achieves state-of-the-art results on the four histopathological image datasets that we have used in this study. We expect it to generalize beyond these four common datasets.

In **Chapter 4**, we investigate the use of state-of-the-art deep learning models, i.e., SqueezeNet, MobileNet, ResNet, and DenseNet, to achieve high classification accuracy. We show that two different approaches, either generating features from pre-trained models (i.e., without fine-tuning), or fine-tuning the CNN from pre-trained models are effective strategies. Our experimental evaluation shows that our classification methods provide better performance in comparison to previously published state-of-art methods. In training an SVM on deep features, we apply ILFS to yield a reduced sub-space of features while maintaining high accuracy. We are able to overcome the significant limitation of deep features which is their high dimensionality.

We find that CNN-based features have the advantage of requiring less time to train than fine-tuning a CNN but require feature extraction and reduction. Alternatively, a CNN model can be fine-tuned end-to-end and used as a classifier, but this process requires more time to train and more computational resources.

Finally, in **Chapter 5**, we have selected a different challenging medical dataset that is captured with MRI in radiology. We have extended our proposed hand-crafted and deep learning techniques to be used on MRI scans. The MRI dataset presents different levels of complexity, i.e., a large number of slices and pattern variability of different slices. Necessary changes were incorporated into the application of those two techniques to handle the MRI

scans efficiently. Consequently, shearlet techniques showed a moderate level of success, but deep learning strategies were adjusted and state-of-the-art results are obtained. When both shearlet-based descriptors and deep features are combined, we have obtained classification results surpassing the state-of-the-art.

6.2 Limitations and Future Work

Although we have delivered effective classification models, there are some limitations of the proposed work. Notably, the classification models that are built for histopathological image classification based on cross-validation protocol, without the use of a hidden set for testing the generalization of our technique. This is a commonly used procedure in the literature due to the fact of limited data points in the medical datasets. Also, it should be acknowledged that building an effective classification model for Alzheimer’s disease comes with the cost of multiple pipelines, i.e., 3D shearlet-based descriptors and deep features.

For future work, in our traditional learning techniques, we utilized the 2D shearlet transform to extract descriptors from histopathological images. We have used PCA for dimensionality reduction, but we plan to investigate other feature reduction methodologies to aggressively reduce the number of descriptors without compromising the classification performance. The problem of a highly imbalanced dataset observed in the BreakHis dataset can be eliminated or reduced by using sampling methods, e.g., Synthetic minority oversampling (SMOTE) technique [34], on the training dataset. Furthermore, we also plan to investigate to resolve the issue of imbalanced datasets in the context of deep learning by utilizing various data augmentation techniques. However, Chen et al. [36] indicated that utilizing unsupervised contrastive learning can gain more from robust data augmentation than supervised learning. Therefore, an intriguing technique worth exploring is unsupervised contrastive learning with data augmentation [82, 36] for small medical datasets. Another direction for future work is to examine alternative weight initialization of deep learning models that were trained on larger datasets than ImageNet. Such deep learning models [107] can be investigated while using a small number of datapoint per class.

Both of our traditional and deep learning techniques have been extended to Alzheimer’s disease classification from MRI data. It would be interesting to examine our approach on another dataset of similar nature of complexity, for example, the Parkinson’s Progression Markers Initiative (PPMI) data repository [138]. Although our techniques were examined for the classification task, it would be interesting to extend our shearlet descriptors and deep learning strategies for segmentation tasks. For example, the multimodal Brain Tumor Image Segmentation Benchmark (BRATS)* is worth exploring. Furthermore, several studies have proposed to build classification models based on selected slices [233, 141], e.g., discarding some slices based on the image entropy. In the future, we plan to investigate a suitable method for slice selection. After selecting a subset of slices that contains the most disease-indicative information, it would be interesting to reduce the complexity of the 3D

*The BRATS benchmark can be accessed via <http://www.brain tumor segmentation.org/>.

problem into 2D. Hence, we can then investigate our texture-based descriptors in the 2D shearlet domain. An alternative approach that can be investigated is the adoption of a more advanced model, e.g., the Squeeze-and-Excitation [88] model while utilizing a channel attention mechanism. In this manner, the model automatically can capture discriminative attributes by combining spatial and channel-wise information.

APPENDICES

Appendix A

Supplemental Material For Chapter 2

In this appendix, we provide a summary of methods that exist in the literature that propose either classical or deep learning techniques.

Table A.1: Summary of methods using BreakHis, Warwick-QU, USCB, Epistroma, or Kather (tumor & strom tissue only).

Authors	Descriptors/Methodology	CL	Dataset	ACC (%)
Reported accuracy of methods using traditional texture analysis				
Kather et al. [103]	A various number of hand-crafted features are extracted: lower-order and higher-order histogram features, LBP, GLCM, GF, Perception-like features. The authors combine some subsets of descriptors to boost the classification result.	SVM	Kather et al.	98.6%
Ribeiro et al. [177]	The authors extract fractal measures and Haralick features from both spatial images and curvelet coefficients. A subset of the features is selected using RELIEFF [110].	PL	Kather et al. (250 examples of each class only)	97.68%
Bianconi et al. [21]	Feature extraction is based on visual perception (i.e., coarseness, contrast, directionality, line-likeness, and roughness).	SVM	Epistroma	96.94%

Continues on the next page

Authors	Descriptors/Methodology	CL	Dataset	ACC (%)
Alinsaif and Lang [5]	A combination of texture features (i.e., GLCM, LBP, LOSIB, SFTA) in the shearlet domain are extracted from both magnitude and phase coefficients and then reduced using PCA.	SVM	Epistroma	98.04%
Smeraldi et al. [197]	A modified version of the texture spectrum technique is proposed to work with partial order rank feature in color space.	1-NN	Epistroma	92.4%
Nava et al. [154]	The authors propose studying histopathological images based on discrete Tchebichef moments.	SVM	Epistroma	96.53%
Ramalho et al. [175]	The authors propose a rotation-invariant feature extraction method based on a structural approach using co-occurrence statistics.	SVM	Epistroma	95%
Song et al. [202]	The authors propose using Fisher vectors with component selective encoding for dimensionality reduction.	SVM	BreaKHis	* PaL: 88.5% - 90.8%; IL: 85% - 87.5%
Bianconi et al. [22]	A modified version of LBP (called IOCLBP) is proposed. Then, IOCLBP is combined with LCC.	NN	BreaKHis	* IL: 91.9% - 93.8%
Ribeiro et al. [177]	The authors extract fractal measures and Haralick features from both spatial images and curvelet coefficients.	PL	Warwick-QU	AUC: 0.994
Dos Santos et al. [52]	The authors propose a technique that combines multiscale, multidimensional, and fuzzy sample entropy strategies. The multiscale strategy is based on various window sizes. The multidimensional strategy is computed for each pixel in the color image. Finally, the goal of the fuzzy strategy is to verify the relevance of the distances between windows.	LR	Warwick-QU	AUC: 0.983

Continues on the next page

Authors	Descriptors/Methodology	CL	Dataset	ACC (%)
Taino et al. [210]	In this study, the authors have proposed a method based on a Genetic algorithm coupled with handcrafted features (i.e., fractal techniques, Haralick texture features, and curvelet coefficients).	RF	Warwick-QU	90.82%
Bruno et al. [28]	LBP is applied on curvelet coefficients. Then, ANOVA feature selection is used to reduce the dimensionality.	PL	UCSB	100%
Yu et al. [235]	In this study, a fusion of two sets of features is obtained: 1) Nuclei segmentation is applied, and then the segmented image is used to extract morphological and spatial features (i.e., Voronoi, Delaunay, and minimum spanning tree), and 2) color deconvolution is applied, and then from each of the decomposed images (i.e., H & E) texture features (SURF, GLCM, LBP) are extracted. Finally, feature vectors are combined and reduced using relief [130].	SVM	UCSB	96.7%
Reported accuracy of deep learning techniques				
Cascianelli et al. [32]	Deep features (i.e., VGG-VD-16 [195]) is used as a feature extractor) with dimensionality reduction.	NN	Epistroma	95.2%
Huang et al. [90]	The authors have proposed unsupervised domain adaptation using transfer learning. A CNN model is trained using different datasets, and cross-validated with an epithelium-stroma dataset.	CNN	Epistroma	93.5%
Bianconi et al. [22]	The authors have proposed using CNN-based (i.e., ResNet-50-FC) features.	NN	Epistroma	97.3%
Wei et al. [226]	Transfer learning of a DL model with data augmentation.	Bi-CNN	BreaKHis	*PaL: 97.02% - 97.89%; IL: 97.56% - 97.97%

Continues on the next page

Authors	Descriptors/Methodology	CL	Dataset	ACC (%)
Nahid et al. [146]	Raw images, hand-crafted features, and frequency domain information are used for breast cancer histopathological image classification.	CNN	BreaKHis	* IL: 94.40% - 97.19%
Jonnalagedda et al. [101]	A deep learning model is trained using augmented images and information from tumor nuclei. All magnification factors are treated as one single dataset.	CNN	BreaKHis	92.2%
Sharma and Mehra [192]	The four magnification factors of the BreaKHis dataset are combined as one dataset. Then, deep features are extracted from a fine-tuned pre-trained VGG16 model.	LR	BreaKHis	92.60%
Alom et al. [11]	The authors of this study propose using the IRRCNN model. This model is trained with data augmentation.	CNN	BreaKHis	* PaL: 97.56% - 97.65%; IL: 97.32% - 97.95%
Vo et al. [221]	First, stain normalization is applied, and Inception-ResNet-v2 is fine-tuned with data augmentation. Various CNN models are trained with resized images (i.e., 600, 450, 300). Finally, deep features are extracted to train a CML model and majority voting is utilized to obtain the final results.	GBT	BreaKHis	* IL: 93.8% - 95.1%
Xie et al. [230]	The Inception ResNet V2 is fine-tuned with data augmentation.	CNN	BreaKHis	* PaL: 99.79% - 100.0%; IL: 99.10% - 99.79%
Jiang et al. [99]	The authors have proposed to used SE-ResNet [88], which the building block of such a model based on the combination of the Squeeze-and-Excitation block.	CNN	BreaKHis	* IL: 99.34% - 98.87%

Continues on the next page

Authors	Descriptors/Methodology	CL	Dataset	ACC (%)
Kumar et al. [113]	First, the dataset is augmented and normalized. Then, a feature vector for each image is obtained using VGG-16, such that a concatenation of descriptors are obtained from multiple layers of the CNN model.	CNN	BreaKHis	* IL: 94.11% - 97.01%
Xu et al. [232]	First, a decision DL model is used. As such, the model aims to decide whether to crop the original image and whether the cropped patch is necessary for the classification stage. Then, the selected cropped patches are used to train another model for the classification, which in turn provides feedback to the decision network.	CNN	BreaKHis	* IL: 97.8% - 98.5%
BenTaieb et al. [19]	Weighted multi-loss terms are proposed for histopathological image classification.	AE	Warwick-QU	95%
Song et al. [200]	The authors in this study have proposed to extract the following features: Dense SIFT, multiple layers of RBM, features of fine-tuned CNN (i.e., VGG-VD). Then, those features are encoded using FV. Finally, a separation-guided technique used to reduce the dimensionality. Particularly, CNN-based encoded FV gave the best accuracy.	SVM	UCSB	98.3%

* In the case of BreaKHis, the range of accuracies indicate the lowest and the highest values achieved when a corresponding technique is examined on the four magnification factors.

Table A.2: Summary of methods using multi-class Kather’s dataset.

Authors	Descriptors/Methodology	CL	ACC (%)
Reported accuracy of methods using traditional texture analysis			
Bianconi et al. [22]	In this study, the authors have proposed to combine IO-CLBP with LCC.	NN	93.4%
Reported accuracy of deep learning techniques			
Wang et al. [222]	Histopathological images are decomposed into H & E stain components.	Bi-CNN	92.6%
Maguolo et al. [136]	The authors propose to fine-tune a pre-trained VGG-16 model. The VGG-16 model is independently retrained each time with a different activation function, and the ensemble of outputs of these DL models is utilized (i.e., sum rule).	CNN	96.88%

Continues on the next page

Authors	Descriptors/Methodology	CL	ACC (%)
Alinsaif & Lang [6]	The authors have utilized DenseNet-201 as an unsupervised feature extractor. To overcome the problem of high-dimensionality, the feature vectors are reduced.	SVM	95.40%
Lai and Chang [118]	Proposed modifying ResNet-50 by replacing the layer before the classification layer (i.e., global average pooling layer) with a flatten layer.	CNN	94.4%
Medela et al. [140]	A DL model is trained on a proposed loss function called a constellation. This proposed loss combines both triplet and multiclass-N-pain loss functions.	CNN	92.96%

Table A.3: Work summary of methods using BACH dataset.

Authors	Descriptors/Methodology	CL	ACC (%)
Reported accuracy of methods using traditional texture analysis			
Sarmiento et al. [187]	Two types of attributes are computed and concatenated for breast cancer image classification: color- and texture-based descriptors.	SVM	79.2%
Reported accuracy of deep learning techniques			
Brancati et al. [25]	Images are down-sampled and a square patch is taken from each image to fine-tune 3 ResNet configurations (i.e., 34, 50, and 101 layers). The outputs of the three networks are combined using the max probability rule to obtain the final label.	CNN	97.3%
Chennamsetty et al. [37]	1) DenseNet-161 and ResNet-101 are fine-tuned using a stain normalization scheme (i.e., mean and standard deviation normalization stats are computed from ImageNet data); 2) DenseNet-161 is fine-tuned based on a different stain normalization scheme (i.e., normalization stats are computed from the whole histological training data). The outputs of (1) & (2) are combined using majority voting.	CNN	97.5%
Kone et al. [109]	The authors propose to Fine-tune ResNeXt50 [231] with data augmentation. First, a histological image is classified into carcinoma & non-carcinoma. If an image is classified as carcinoma, then another binary classifier is used to find the specific label (i.e., in situ or invasive). Similarly, the same approach is applied when an image is classified as non-carcinoma (i.e. normal or benign).	CNN	99%

Continues on the next page

Authors	Descriptors/Methodology	CL	ACC (%)
Vesal et al. [219]	Inception-V3 is fine-tuned based on color-corrected image patches (i.e., color histogram stats are matched between source and target image) and augmented data.	CNN	97.08%
Nazeri et al. [155]	A model is trained based on color-augmented patches of histological images. Feature maps of all patches of an image are stacked and form feature vectors for all images. Then, another model is trained to make the final prediction.	CNN	95%
Wang et al. [224]	Data augmentation (including image deformation) is applied to fine-tune VGG16.	CNN	91.7%
Pimkin et al. [167]	Two different models (i.e., ResNet and DenseNet) are fine-tuned, which are fine-tuned based on patches that are extracted from the histological images. From those fine-tuned models different features are extracted (e.g., min, max, or mean values) to be fed to a CML.	XG-Boost	90%
Vo et al. [221]	First, stain normalization is applied, and Inception-ResNet-v2 is fine-tuned with data augmentation. Various CNN models are trained with resized images (i.e., 600, 450, 300). Finally, deep features are extracted to train a CML model and majority voting is utilized to obtain the final results.	GBT	96.4%
Minh et al. [120]	The authors propose to fine-tune a modified Inception-V3. The output feature maps from block-9, block-10, and block-11 are concatenated prior to being fed to a global average pooling layer.	CNN	95%

Table A.4: Summary of methods using HistologyDS dataset.

Authors	Descriptors/Methodology	CL	ACC (%)
Reported accuracy of methods using traditional texture analysis			
Lumini et al. [134]	Images are pre-processed to obtain wavelets, gradient images, a Gaussian image pyramid, and represented in different color spaces (i.e., RGB, HSV, and YUV). All stacked together to form an n-layer image. The following features are extracted: from gray-level (LPQ & LBP) and from the n-layer image (Volumetric LBP, Volumetric LPQ, and LPQ-TOP). However, the weighted sum rule of two trained models using LPQ and LPQ-TOP gave the best AUC.	SVM	AUC: 0.936

Continues on the next page

Authors	Descriptors/Methodology	CL	ACC (%)
Di Ru- berto et al. [47]	Three sets of features are extracted: GLCM features from the red channel, GLRLM features from the green channel, GLDM features from the blue channel. A feature selection technique is applied for each set of features, and then combined and fed to train a CML model.	SVM	96.0%
Dos San- tos et al. [51]	Texture descriptors are extracted from edge regions in the histological image. The Sobel edge detector or Difference of Gaussian is applied to obtain an edge and non-edge maps. Edge and non-edge maps are fed to LTP an LPQ for feature extraction. Thereafter, the sum rule fusion of different trained models is used.	SVM	AUC: 0.929
Graña et al. [72]	A histological image is converted into a graph by connecting each pixel to its 4-neighbors. From this graph, the Dijkstra's cost scores are compared with the pixels' geometric distance to obtain the minimum cost path. From the gray-level image, LTP features are also extracted. A final label is assigned based on the weighted sum rule.	SVM	AUC: 0.929
Nanni et al. [151]	A number of descriptors are extracted: RICLBP, Ensemble of Multi-ternary coding version of Multi-scale LPQ, multi-quinary coding version of Multi-scale linear configuration, and BSIF. CML models are trained based on the aforementioned features and fused using the sum rule.	SVM	AUC: 0.948
Reported accuracy of methods using deep learning techniques			
Nanni et al. [150]	The following descriptors are extracted: Deep features based on different CNN models, and Loci which use an ensemble of LQP to extract geometric loci of points: circle, ellipse, parabola, and archimedean spiral. To obtain the final label, the outputs are fused by the weighted sum rule.	SVM	AUC: 0.941
Lai and Deng [119]	Deep features are extracted from a DL model trained from scratch. Traditional features, i.e., color moments and GLCM features, are also extracted. The two sets of features are combined to train a CML model.	MLP	90.2%

Table A.5: Summary of methods using Lymphoma dataset.

Authors	Descriptors/Methodology	CL	ACC (%)
Reported accuracy of methods using traditional texture analysis			
Mohan et al. [144]	The authors have utilized a tool called Wndchrm [190] to extract features from an image. Then, a Fisher score technique is used for feature selection [73].	MD	94.78%
Do Nascimento et al. [49]	Histological images are pre-processed for color enhancement using CLAHE and 2D order-statistics. Then, a global threshold is used to identify the regions of interest to extract morphological features and non-morphological features. Ansari-Bradley feature selection is applied to reduce the dimensionality. Lymphoma dataset is a 3-class classification problem, but the authors have applied the following binary classification problems: CLL-MCL, CLL-FL, and FL-MCL.	PL	100%
Roberto et al. [178]	A histological image is divided into smaller blocks, and then the color similarity of pixels is computed. The smaller regions of interest are labeled using a clustering technique to extract the following features: the number of clusters, the occurrence of percolation, and the coverage ratio of the largest cluster.	DECO-AUC: RATE	0.943
Gonzalez et al. [71]	DSMs are extracted in sliding window fashion from an image. N-way ANOVA is used to check the statistical significance of the descriptors and window size.	RF	100%
Nava et al. [153]	The following discrete orthogonal moments are extracted: moments of Discrete Tchebichef, Discrete Krawtchouk, and Discrete Dual Hahn. The aforementioned moments are extracted from H & E components separately using a sliding window. The dimension of features is reduced using RELIEFF [110].	KFDA	93.83%
Song et al. [201]	The authors have explored the following features: 1) Improved FV [166], 2) LBP, 3) histogram of oriented gradients [60], 4) CENTRIST and, 5) GIST texture features [159]. The authors combine and transform the descriptors using subcategory discriminant transform. Such a transform is used to learn convolution kernels and these kernels used to transform the feature vectors which then fed to train a CML.	SVM	96.8%

Continues on the next page

Authors	Descriptors/Methodology	CL	ACC (%)
Orlov et al. [161]	A color deconvolution is applied to obtained H & E components. Particularly, the extracted features from the eosin components yielded the best accuracy. The eosin components are transformed into the following domains for feature extraction: wavelets, Chebyshev, and Fourier. Then features are computed and combined to form a single raw vector of a histological image.	RBF	99%
Reported accuracy of methods using deep learning techniques			
Nanni et al. [152]	Different DL models are fine-tuned for deep feature extraction (i.e., GoogLeNet, VGGNet, ResNet, Inception, and Inception-ResNet-v2). Also, different hand-crafted features are extracted (i.e., LTP, CLBP, Multiscale RICLBP, LETRIST [199], morphological features, AHP, color descriptors [24], Ensemble of LPQ by changing the filter size, and BSIF). Then, CML models are trained based on these deep features and hand-crafted descriptors. The final label is predicted based on the fusion of the two scores using the sum rule.	SVM	97.33%
Song et al. [200]	The following features are extracted: Dense SIFT, multi-layers descriptors from RBM, descriptors from fine-tuned CNNs (i.e., VGG-VD). Then, those features are encoded using FV. Finally, a separation-guided technique is used to reduce the dimensionality.	SVM	97.9%

Table A.6: Summary of methods using Path960 dataset.

Authors	Descriptors/Methodology	CL	ACC (%)
Reported accuracy of methods using traditional texture analysis			
Kumar et al. [114]	A comparative study of three techniques is conducted: LBP, bag-of-visual words, and deep features. Bag-of-visual words achieved the highest classification accuracy.	SVM	96.50%
Alhindi [4]	LBP, the histogram of gradients, and a pre-trained deep neural network are used to classify the dataset. The authors have found that LBP achieves higher accuracy than deep features.	SVM	90.52%
Sukhia et al. [208]	In this study, histopathological image retrieval is based on multi-scale, the multi-channel decoded local ternary pattern features followed by a vector of locally aggregated descriptors.	L2 distance	Precision: 95%

Continues on the next page

Authors	Descriptors/Methodology	CL	ACC (%)
Liu et al. [129]	The authors propose to combine LBP variant and bag-of-visual-words descriptors with deep features.	SVM	93.49%
Reported accuracy of methods using deep learning techniques			
Alom et al. [10]	The authors propose to use the IRRCNN with different data augmentation techniques (i.e., rotation, shear, zooming, and horizontal and vertical flipping).	CNN	98.79%

Appendix B

Supplemental Material For Chapter 3

In this appendix, we provide the classification performance of our methods of Chapter 3 in the form of tables.

Table B.1: Classification performance on Kather’s dataset using shearlet-based descriptors.

		RP + Magnitude		RP		Magnitude	
Method	P	SVM	DTB	SVM	DTB	SVM	DTB
(1): Baseline performance of existing shearlet-based methods							
Vo et al.	ACC%	82.34% ± 1.87%	79.42%± 1.27%	78.46%± 2.11%	75.22%± 2.17%	71.10%± 1.65%	68.28%± 1.09%
	AUC	0.9658 ± 0.0038	0.9646 ± 0.0038	0.9545 ± 0.0049	0.9552 ± 0.0066	0.9283 ± 0.0092	0.9340 ± 0.0036
	Sen	0.8234 ± 0.0186	0.7943 ± 0.0127	0.7846 ± 0.0212	0.7522 ± 0.0218	0.7111 ± 0.0161	0.6828 ± 0.0107
	Prec	0.8263 ± 0.0191	0.7961 ± 0.0131	0.7858 ± 0.0218	0.7568 ± 0.0213	0.7102 ± 0.0180	0.6826 ± 0.0131
Meshkini and Ghassemian	ACC%	<i>80.06%</i> ± <i>1.54%</i>	<i>83.06%</i> ± <i>1.33%</i>	<i>37.72%</i> ± <i>1.78%</i>	<i>77.84%</i> ± <i>1.17%</i>	<i>83.38%</i> ± <i>1.81%</i>	<i>83.74%</i> ± <i>1.79%</i>
	AUC	<i>0.9648</i> ± <i>0.0055</i>	<i>0.9800</i> ± <i>0.0081</i>	<i>0.7553</i> ± <i>0.0115</i>	<i>0.9649</i> ± <i>0.0028</i>	<i>0.9738</i> ± <i>0.0036</i>	<i>0.9816</i> ± <i>0.0027</i>
	Sen	<i>0.8006</i> ± <i>0.0154</i>	<i>0.8307</i> ± <i>0.0134</i>	<i>0.3772</i> ± <i>0.0179</i>	<i>0.7784</i> ± <i>0.0117</i>	<i>0.8338</i> ± <i>0.0181</i>	<i>0.8374</i> ± <i>0.0179</i>
	Prec	<i>0.8111</i> ± <i>0.0128</i>	<i>0.8325</i> ± <i>0.0124</i>	<i>0.4338</i> ± <i>0.0193</i>	<i>0.7792</i> ± <i>0.0105</i>	<i>0.8385</i> ± <i>0.0170</i>	<i>0.8400</i> ± <i>0.0168</i>
Zhou et al.	ACC%	<i>64.26%</i> ± <i>2.20%</i>	<i>67.78%</i> ± <i>1.38%</i>	<i>51.34%</i> ± <i>2.35%</i>	<i>54.68%</i> ± <i>1.79%</i>	<i>60.06%</i> ± <i>1.43%</i>	<i>64.42%</i> ± <i>1.82%</i>
	AUC	<i>0.9016</i> ± <i>0.0074</i>	<i>0.9337</i> ± <i>0.0051</i>	<i>0.8519</i> ± <i>0.0104</i>	<i>0.8846</i> ± <i>0.0045</i>	<i>0.8764</i> ± <i>0.0050</i>	<i>0.9132</i> ± <i>0.0061</i>

Continues on the next page

Table B.1: Classification performance on Kather’s dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
Dong et al.	Sen	0.6425 ± 0.0222	0.6778 ± 0.0138	0.5134 ± 0.0233	0.5468 ± 0.0177	0.6006 ± 0.0143	0.6442 ± 0.0183
	Prec	0.6494 ± 0.0240	0.6842 ± 0.0208	0.5112 ± 0.0235	0.5540 ± 0.0210	0.6182 ± 0.0198	0.6511 ± 0.0197
	ACC%	$81.10\% \pm 1.90\%$	$77.50\% \pm 1.17\%$	$70.66\% \pm 1.81\%$	$72.10\% \pm 0.97\%$	$76.62\% \pm 1.86\%$	$72.08\% \pm 1.65\%$
	AUC	0.9621 ± 0.0046	0.9620 ± 0.0033	0.9297 ± 0.0082	0.9457 ± 0.0037	0.9504 ± 0.0063	0.9448 ± 0.0046
	Sen	0.8110 ± 0.0188	0.7751 ± 0.0116	0.7066 ± 0.0179	0.7210 ± 0.0100	0.7662 ± 0.0185	0.7207 ± 0.0166
	Prec	0.8146 ± 0.0187	0.7753 ± 0.0140	0.7114 ± 0.0187	0.7263 ± 0.0106	0.7681 ± 0.0190	0.7210 ± 0.0172
	(2): Proposed shearlet-based methods for textured bio-medical image classification						
	ACC%	$86.24\% \pm 1.27\%$	$80.98\% \pm 1.54\%$	$82.82\% \pm 1.47\%$	$78.48\% \pm 1.30\%$	$80.28\% \pm 1.64\%$	$76.30\% \pm 1.16\%$
	AUC	0.9761 ± 0.0027	0.9712 ± 0.0031	0.9697 ± 0.0035	0.9636 ± 0.0046	0.9603 ± 0.0039	0.9568 ± 0.0051
CM	Sen	0.8624 ± 0.0128	0.8098 ± 0.0155	0.8282 ± 0.0148	0.7849 ± 0.01274	0.8028 ± 0.0166	0.7630 ± 0.0114
	Prec	0.8649 ± 0.0130	0.8127 ± 0.0141	0.8322 ± 0.0147	0.7884 ± 0.0115	0.8037 ± 0.0157	0.7642 ± 0.0127
	ACC%	$84.24\% \pm 0.91\%$	$77.90\% \pm 1.75\%$	$81.36\% \pm 1.16\%$	$75.10\% \pm 2.20\%$	$79.36\% \pm 1.39\%$	$75.38\% \pm 1.71\%$
	AUC	0.9724 ± 0.0026	0.9630 ± 0.0043	0.9649 ± 0.0039	0.9559 ± 0.0053	0.9590 ± 0.0037	0.9541 ± 0.0064
LBP	Sen	0.8424 ± 0.0092	0.7791 ± 0.0174	0.8136 ± 0.0115	0.7511 ± 0.0216	0.7936 ± 0.0140	0.7537 ± 0.0172
	Prec	0.8454 ± 0.0106	0.7829 ± 0.0185	0.8171 ± 0.0129	0.7551 ± 0.0231	0.7964 ± 0.0131	0.7573 ± 0.0164
	ACC%	$86.32\% \pm 1.36\%$	$79.78\% \pm 2.11\%$	$82.36\% \pm 1.82\%$	$79.76\% \pm 1.35\%$	$76.52\% \pm 1.69\%$	$71.24\% \pm 1.01\%$
	AUC	0.9773 ± 0.0021	0.9703 ± 0.0046	0.9671 ± 0.0045	0.9689 ± 0.0024	0.9495 ± 0.0048	0.9402 ± 0.0037
LOSIB	Sen	0.8632 ± 0.0134	0.7978 ± 0.0209	0.8236 ± 0.0180	0.7976 ± 0.0136	0.7653 ± 0.0169	0.7124 ± 0.0098
	Prec	0.8664 ± 0.0128	0.8017 ± 0.0216	0.8274 ± 0.0169	0.8008 ± 0.0127	0.7672 ± 0.0191	0.7128 ± 0.0123
	ACC%	$82.92\% \pm 1.41\%$	$78.24\% \pm 1.21\%$	$78.10\% \pm 1.43\%$	$75.04\% \pm 1.44\%$	$79.72\% \pm 1.07\%$	$75.42\% \pm 2.23\%$
	SFTA						

Continues on the next page

Table B.1: Classification performance on Kather’s dataset using shearlet-based descriptors.

		RP + Magnitude		RP		Magnitude	
Method	P	SVM	DTB	SVM	DTB	SVM	DTB
	AUC	0.9682 $\pm$	0.9645 $\pm$	0.9554 $\pm$	0.9570 $\pm$	0.9597 $\pm$	0.9574 $\pm$
		0.0030	0.0035	0.0031	0.0047	0.0032	0.0054
	Sen	0.8292 $\pm$	0.7825 $\pm$	0.7810 $\pm$	0.7504 $\pm$	0.7972 $\pm$	0.7543 $\pm$
		0.0143	0.0124	0.0141	0.0144	0.0110	0.0224
	Prec	0.8331 $\pm$	0.7845 $\pm$	0.7877 $\pm$	0.7539 $\pm$	0.7986 $\pm$	0.7587 $\pm$
		0.0139	0.0117	0.0144	0.0133	0.0109	0.0248
(3): Integrating shearlet-based existing techniques with our proposed methods							
Fusion #1	ACC%	87.54% $\pm$	81.40% $\pm$	85.32% $\pm$	80.30% $\pm$	83.92% $\pm$	77.76% $\pm$
		1.85%	1.21%	1.72%	1.52%	1.84%	1.21%
	AUC	0.9823 $\pm$	0.9734 $\pm$	0.9774 $\pm$	0.9707 $\pm$	0.9740 $\pm$	0.9619 $\pm$
		0.0042	0.0019	0.0043	0.0024	0.0049	0.0031
	Sen	0.8755 $\pm$	0.8141 $\pm$	0.8533 $\pm$	0.8029 $\pm$	0.8392 $\pm$	0.7776 $\pm$
Fusion #2	ACC%	88.46% $\pm$	82.28% $\pm$	85.14% $\pm$	80.26% $\pm$	83.54% $\pm$	77.20% $\pm$
		1.14%	1.46%	1.05%	2.11%	1.30%	1.92%
	AUC	0.9804 $\pm$	0.9746 $\pm$	0.9749 $\pm$	0.9699 $\pm$	0.9693 $\pm$	0.9606 $\pm$
		0.0023	0.0034	0.0029	0.0049	0.0033	0.0040
	Sen	0.8846 $\pm$	0.8228 $\pm$	0.8514 $\pm$	0.8026 $\pm$	0.8354 $\pm$	0.7720 $\pm$
Fusion #3	ACC%	88.74% $\pm$	82.53% $\pm$	85.51% $\pm$	80.70% $\pm$	83.75% $\pm$	77.44% $\pm$
		0.0108	0.0149	0.0107	0.0198	0.0130	0.0182
	AUC	92.54% $\pm$	88.60% $\pm$	-	-	-	-
		1.32%	2.03%	-	-	-	-
	Sen	0.9906 $\pm$	0.9900 $\pm$	-	-	-	-
Fusion #3	ACC%	0.9254 $\pm$	0.8860 $\pm$	-	-	-	-
		0.0132	0.0203	-	-	-	-
	AUC	0.9267 $\pm$	0.8881 $\pm$	-	-	-	-
		0.0129	0.0194	-	-	-	-

Table B.2: Classification performance on BreakHis dataset using shearlet-based descriptors.

		RP + Magnitude		RP		Magnitude	
Method	P	SVM	DTB	SVM	DTB	SVM	DTB
(1): Baseline performance of existing shearlet-based methods							
Vo et al.	ACC%	85.69%±1.18%	80.65%±1.05%	83.63%±0.99%	78.95%±0.74%	76.10%±0.66%	76.52%±0.64%
	AUC	0.9121±0.0053	0.8687 ±0.0079	0.8915 ±0.0103	0.8443 ±0.0183	0.7717 ±0.0123	0.7926 ±0.0175
	Sen	0.7210±0.0276	0.4855 ±0.0346	0.6766 ±0.0231	0.4496 ±0.0157	0.4698 ±0.0213	0.4056 ±0.0301
	Prec	0.8035±0.0301	0.8268 ±0.0296	0.7732 ±0.0201	0.7881 ±0.0208	0.6701 ±0.0192	0.7247 ±0.0159
Meshkini and Ghassemian	ACC%	78.06%±0.80%	79.43%±1.09%	77.33%±1.23%	79.76%±0.71%	76.44%±0.96%	77.46%±0.85%
	AUC	0.8339±0.0030	0.8523±0.0055	0.8111±0.0086	0.8579±0.0078	0.7991 ±0.0139	0.8357 ±0.0112
	Sen	0.5177±0.0194	0.4927±0.0213	0.4899±0.0291	0.4980±0.0237	0.4722 ±0.0232	0.4218 ±0.0246
	Prec	0.7049±0.0223	0.7688±0.0313	0.6973±0.0284	0.7779±0.0300	0.6788 ±0.0215	0.7496 ±0.0181
Zhou et al.	ACC%	68.66%±1.23%	70.21%±0.36%	65.20%±1.47%	69.64%±0.47%	66.47%±0.88%	70.58%±0.49%
	AUC	0.6426±0.0245	0.6781±0.0173	0.5832±0.0193	0.6154±0.0116	0.6022 ±0.0198	0.6709 ±0.0165
	Sen	0.2875±0.0310	0.0677±0.0139	0.2290±0.0273	0.0456±0.0096	0.2415 ±0.0326	0.0956 ±0.0123
	Prec	0.4996±0.0357	0.7915±0.0314	0.4035±0.0391	0.7728±0.1106	0.4361 ±0.0291	0.7436 ±0.0687
Dong et al.	ACC%	79.49%±1.39%	78.40%±0.54%	73.54%±1.61%	75.58%±0.78%	78.40 ±1.20%	77.56%±0.97%
	AUC	0.8390±0.0163	0.8359±0.0110	0.7641±0.0172	0.8018±0.0069	0.8090 ±0.0114	0.8082 ±0.0156
	Sen	0.5980±0.0342	0.4360±0.0139	0.4860±0.0312	0.3328±0.0327	0.5176 ±0.0270	0.4380 ±0.0129
	Prec	0.7077±0.0226	0.7855±0.0169	0.6011±0.0324	0.7600±0.0155	0.7204 ±0.0243	0.7480 ±0.0307
(2): Proposed shearlet-based methods for textured bio-medical image classification							
CM	ACC%	87.26%±1.18%	79.33%±1.17%	86.93%±0.75%	79.95%±0.39%	78.85%±1.42%	75.08%±0.80%

Continues on the next page

Table B.2: Classification performance on BreakHis dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
	AUC	0.9365 $\pm$ 0.0104	0.8679 $\pm$ 0.0134	0.9283 $\pm$ 0.0052	0.8687 $\pm$ 0.0085	0.8350 $\pm$ 0.0143	0.7823 $\pm$ 0.0148
	Sen	0.7391 $\pm$ 0.0286	0.4137 $\pm$ 0.0351	0.7520 $\pm$ 0.0242	0.4677 $\pm$ 0.0188	0.5649 $\pm$ 0.0371	0.2911 $\pm$ 0.0208
	Prec	0.8358 $\pm$ 0.0230	0.8500 $\pm$ 0.0270	0.8166 $\pm$ 0.0108	0.8140 $\pm$ 0.0164	0.7021 $\pm$ 0.0253	0.7717 $\pm$ 0.0274
LBP	ACC%	89.51% $\pm$ 0.69%	81.63% $\pm$ 0.96%	87.15% $\pm$ 0.51%	79.58% $\pm$ 1.05%	86.07% $\pm$ 1.04%	80.01% $\pm$ 1.11%
	AUC	0.9477 $\pm$ 0.0043	0.8905 $\pm$ 0.0101	0.9296 $\pm$ 0.0054	0.8665 $\pm$ 0.0131	0.9153 $\pm$ 0.0080	0.8630 $\pm$ 0.0075
	Sen	0.7964 $\pm$ 0.0233	0.5081 $\pm$ 0.0180	0.7508 $\pm$ 0.0173	0.4484 $\pm$ 0.0305	0.7177 $\pm$ 0.0355	0.4641 $\pm$ 0.0296
	Prec	0.8590 $\pm$ 0.0149	0.8444 $\pm$ 0.0299	0.8241 $\pm$ 0.0126	0.8182 $\pm$ 0.0212	0.8158 $\pm$ 0.0132	0.8200 $\pm$ 0.0219
LOSIB	ACC%	87.81% $\pm$ 0.63%	80.63% $\pm$ 1.03%	85.78% $\pm$ 0.52%	78.94% $\pm$ 1.47%	80.31% $\pm$ 1.09%	78.49% $\pm$ 0.90%
	AUC	0.9295 $\pm$ 0.0056	0.8771 $\pm$ 0.0073	0.9092 $\pm$ 0.0042	0.8516 $\pm$ 0.0174	0.8486 $\pm$ 0.0085	0.8206 $\pm$ 0.0119
	Sen	0.7585 $\pm$ 0.0252	0.4972 $\pm$ 0.0225	0.7234 $\pm$ 0.0177	0.4315 $\pm$ 0.0284	0.5665 $\pm$ 0.0274	0.4617 $\pm$ 0.0170
	Prec	0.8380 $\pm$ 0.0163	0.8121 $\pm$ 0.0220	0.8036 $\pm$ 0.0116	0.8071 $\pm$ 0.0436	0.7447 $\pm$ 0.0220	0.7580 $\pm$ 0.0247
SFTA	ACC%	89.72%$\pm$ 0.63%	81.83% $\pm$ 0.79%	87.95% $\pm$ 0.57%	79.90% $\pm$ 1.20%	83.53% $\pm$ 1.04%	80.39% $\pm$ 1.14%
	AUC	0.9527$\pm$ 0.0055	0.8880 $\pm$ 0.0071	0.9393 $\pm$ 0.0065	0.8769 $\pm$ 0.0070	0.8943 $\pm$ 0.0090	0.8424 $\pm$ 0.0158
	Sen	0.8040$\pm$ 0.0243	0.4992 $\pm$ 0.0265	0.7701 $\pm$ 0.0239	0.4601 $\pm$ 0.0251	0.6645 $\pm$ 0.0332	0.4665 $\pm$ 0.0220
	Prec	0.8593$\pm$ 0.0112	0.8641 $\pm$ 0.0137	0.8332 $\pm$ 0.0112	0.8197 $\pm$ 0.0324	0.7787 $\pm$ 0.0264	0.8355 $\pm$ 0.0323
(3): Integrating shearlet-based existing techniques with our proposed methods							
Fusion #1	ACC%	91.28%$\pm$ 0.51%	81.78% $\pm$ 0.73%	89.58% $\pm$ 0.83%	80.98% $\pm$ 0.71%	87.29% $\pm$ 0.37%	80.30% $\pm$ 0.62%
	AUC	0.9650$\pm$ 0.0031	0.8981 $\pm$ 0.0053	0.9515 $\pm$ 0.0031	0.8853 $\pm$ 0.0102	0.9346 $\pm$ 0.0051	0.8719 $\pm$ 0.0084
	Sen	0.8391$\pm$ 0.0147	0.5085 $\pm$ 0.0087	0.8121 $\pm$ 0.0119	0.4855 $\pm$ 0.0252	0.7561 $\pm$ 0.0214	0.4504 $\pm$ 0.0163

Continues on the next page

Table B.2: Classification performance on BreakHis dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
	Prec	0.8775$\pm$0.0120	0.8508 $\pm$ 0.0259	0.8495 $\pm$ 0.0219	0.8412 $\pm$ 0.0185	0.8244 $\pm$ 0.0094	0.8516 $\pm$ 0.0198
Fusion #2	ACC%	89.33% $\pm$ 0.61%	80.33% $\pm$ 0.89%	87.82% $\pm$ 0.36%	80.24% $\pm$ 0.84%	83.23% $\pm$ 0.93%	77.80% $\pm$ 0.51%
	AUC	0.9503 $\pm$ 0.0053	0.8763 $\pm$ 0.0138	0.9354 $\pm$ 0.0037	0.8742 $\pm$ 0.0072	0.8872 $\pm$ 0.0098	0.8446 $\pm$ 0.0071
	Sen	0.7964 $\pm$ 0.0245	0.4544 $\pm$ 0.0202	0.7690 $\pm$ 0.0132	0.4726 $\pm$ 0.0171	0.6536 $\pm$ 0.0155	0.3694 $\pm$ 0.0173
	Prec	0.8537 $\pm$ 0.0105	0.8471 $\pm$ 0.0213	0.8304 $\pm$ 0.0098	0.8216 $\pm$ 0.0252	0.7770 $\pm$ 0.0256	0.8270 $\pm$ 0.0170
Fusion #3	ACC%	90.10% $\pm$ 0.87%	85.08% $\pm$ 0.80%	-	-	-	-
	AUC	0.9560 $\pm$ 0.0053	0.9256 $\pm$ 0.0065	-	-	-	-
	Sen	0.8028 $\pm$ 0.0113	0.5992 $\pm$ 0.0175	-	-	-	-
	Prec	0.8723 $\pm$ 0.0281	0.8887 $\pm$ 0.0161	-	-	-	-

Table B.3: Classification performance on Epistroma dataset using shearlet-based descriptors.

		RP + Magnitude		RP		Magnitude	
Method	P	SVM	DTB	SVM	DTB	SVM	DTB
(1): Baseline performance of existing shearlet-based methods							
Vo et al.	ACC%	96.44%±1.47%	95.42%±1.03%	95.50%±1.52%	94.26%±1.95%	94.48%±1.54%	94.26%±1.35%
	AUC	0.9895±0.0082	0.9897 ± 0.0047	0.9873 ± 0.0078	0.9839 ± 0.0087	0.9830 ± 0.0083	0.9763 ± 0.0102
	Sen	0.9685±0.0131	0.9443 ± 0.0162	0.9685 ± 0.0117	0.9467 ± 0.0268	0.9588 ± 0.0182	0.9564 ± 0.0152
	Prec	0.9722±0.0170	0.9787 ± 0.0083	0.9577 ± 0.0260	0.9576 ± 0.0220	0.9501 ± 0.0219	0.9487 ± 0.0193
Meshkini and Ghassemian	ACC%	<i>96.08%±1.24%</i>	<i>96.29%±1.86%</i>	<i>95.35%±1.37%</i>	<i>96.37%±2.30%</i>	<i>95.35%±1.49%</i>	<i>96.44%±2.20%</i>
	AUC	<i>0.9926±0.0046</i>	<i>0.9929±0.0041</i>	<i>0.9887±0.0058</i>	<i>0.9919±0.0077</i>	<i>0.9899 ± 0.0050</i>	<i>0.9908 ± 0.0068</i>

Continues on the next page

Table B.3: Classification performance on Epistroma dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
Zhou et al.	Sen	0.9710 ± 0.0190	0.9600 ± 0.0222	0.9685 ± 0.0207	0.9600 ± 0.0269	0.9722 ± 0.0161	0.9564 ± 0.0274
	Prec	0.9642 ± 0.0169	0.9778 ± 0.0151	0.9550 ± 0.0183	0.9791 ± 0.0184	0.9517 ± 0.0186	0.9840 ± 0.0181
	ACC%	$89.68\% \pm 1.33\%$	$88.30\% \pm 2.68\%$	$81.69\% \pm 2.53\%$	$79.65\% \pm 3.11\%$	$84.45\% \pm 2.98\%$	$87.13\% \pm 2.95\%$
	AUC	0.9506 ± 0.0158	0.9400 ± 0.0223	0.8748 ± 0.0222	0.8721 ± 0.0367	0.9110 ± 0.0194	0.9416 ± 0.0244
	Sen	0.9370 ± 0.0219	0.9103 ± 0.0287	0.8934 ± 0.0363	0.8666 ± 0.0343	0.9236 ± 0.0141	0.9102 ± 0.0327
	Prec	0.8961 ± 0.0183	0.8968 ± 0.0275	0.8186 ± 0.0251	0.8088 ± 0.0301	0.8357 ± 0.0330	0.8799 ± 0.0282
	ACC%	$95.72\% \pm 1.54\%$	$95.28\% \pm 1.66\%$	$94.84\% \pm 1.59\%$	$93.39\% \pm 2.11\%$	$95.86\% \pm 1.32\%$	$93.75\% \pm 1.15\%$
	AUC	0.9882 ± 0.0071	0.9837 ± 0.0089	0.9836 ± 0.0071	0.9776 ± 0.0132	0.9884 ± 0.0082	0.9836 ± 0.0101
Dong et al.	Sen	0.9551 ± 0.0162	0.9503 ± 0.0271	0.9503 ± 0.0218	0.9187 ± 0.0300	0.9685 ± 0.0103	0.9600 ± 0.0182
	Prec	0.9735 ± 0.0227	0.9707 ± 0.0178	0.9637 ± 0.0211	0.9697 ± 0.0201	0.9630 ± 0.0194	0.9381 ± 0.0231
(2): Proposed shearlet-based methods for textured bio-medical image classification							
CM	ACC%	$97.24\% \pm 1.27\%$	$94.41\% \pm 1.48\%$	$96.66\% \pm 1.14\%$	$94.99\% \pm 1.43\%$	$96.00\% \pm 1.30\%$	$94.19\% \pm 2.19\%$
	AUC	0.9917 ± 0.0067	0.9819 ± 0.0126	0.9925 ± 0.0064	0.9828 ± 0.0116	0.9863 ± 0.0087	0.9828 ± 0.0091
	Sen	0.9733 ± 0.0126	0.9419 ± 0.0258	0.9672 ± 0.0129	0.9503 ± 0.0134	0.9636 ± 0.0163	0.9418 ± 0.0294
	Prec	0.9807 ± 0.0150	0.9646 ± 0.0202	0.9769 ± 0.0130	0.9658 ± 0.0175	0.9696 ± 0.0116	0.9607 ± 0.0186
LBP	ACC%	$95.64\% \pm 1.32\%$	$95.57\% \pm 2.04\%$	$95.50\% \pm 1.06\%$	$94.04\% \pm 1.56\%$	$95.71\% \pm 1.39\%$	$93.90\% \pm 1.75\%$
	AUC	0.9890 ± 0.0073	0.9905 ± 0.0069	0.9871 ± 0.0076	0.9850 ± 0.0054	0.9876 ± 0.0088	0.9800 ± 0.0113
	Sen	0.9624 ± 0.0146	0.9479 ± 0.0261	0.9661 ± 0.0178	0.9394 ± 0.0221	0.9600 ± 0.0153	0.9406 ± 0.0239
	Prec	0.9653 ± 0.0203	0.9779 ± 0.0210	0.9594 ± 0.0133	0.9607 ± 0.0176	0.9687 ± 0.0206	0.9573 ± 0.0194

Continues on the next page

Table B.3: Classification performance on Epistroma dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
LOSIB	ACC%	96.29%± 1.47%	95.13%± 1.61%	95.50%± 1.32%	96.65%± 1.51%	95.35%± 1.85%	93.10%± 1.29%
	AUC	0.9900 ± 0.0069	0.9870 ± 0.0074	0.9868 ± 0.0073	0.9911 ± 0.0065	0.9892 ± 0.0087	0.9825 ± 0.0103
	Sen	0.9733 ± 0.0096	0.9454 ± 0.0245	0.9721 ± 0.0152	0.9624 ± 0.0218	0.9758 ± 0.0141	0.9503 ± 0.0276
	Prec	0.9656 ± 0.0215	0.9729 ± 0.0179	0.9541 ± 0.0203	0.9816 ± 0.0132	0.9487 ± 0.0239	0.9362 ± 0.0194
SFTA	ACC%	95.71%± 1.54%	93.90%± 1.81%	95.28%± 2.16%	91.57%± 1.91%	94.33%± 1.52%	92.01%± 2.07%
	AUC	0.9881 ± 0.0064	0.9841 ± 0.0106	0.9870 ± 0.0068	0.9775 ± 0.0093	0.9824 ± 0.0080	0.9732 ± 0.0125
	Sen	0.9661 ± 0.0137	0.9297 ± 0.0282	0.9624 ± 0.0145	0.9006 ± 0.0265	0.9636 ± 0.0114	0.9455 ± 0.0221
	Prec	0.9628 ± 0.0169	0.9676 ± 0.0184	0.9593 ± 0.0233	0.9572 ± 0.0266	0.9438 ± 0.024	0.9232 ± 0.0178
(3): Integrating shearlet-based existing techniques with our proposed methods							
Fusion #1	ACC%	96.59%± 1.24%	95.93%± 1.95%	96.37%± 1.14%	94.98%± 0.73%	95.71%± 1.25%	95.13%± 1.37%
	AUC	0.9889 ± 0.0111	0.9910 ± 0.0070	0.9889 ± 0.0103	0.9838 ± 0.0131	0.9877 ± 0.0104	0.9841 ± 0.0104
	Sen	0.9612 ± 0.0126	0.9624 ± 0.0224	0.9612 ± 0.0179	0.9406 ± 0.0194	0.9563 ± 0.0155	0.9503 ± 0.0201
	Prec	0.9816 ± 0.0140	0.9696 ± 0.0175	0.9780 ± 0.0135	0.9752 ± 0.0139	0.9721 ± 0.0185	0.9682 ± 0.0172
Fusion #2	ACC%	97.09%± 1.49%	94.98%± 1.56%	96.37%± 1.45%	95.86%± 2.11%	96.73%± 1.33%	95.28%± 1.54%
	AUC	0.9910 ± 0.0099	0.9829 ± 0.0111	0.9908 ± 0.0092	0.9884 ± 0.0113	0.9892 ± 0.0097	0.9872 ± 0.0106
	Sen	0.9660 ± 0.0205	0.9382 ± 0.0211	0.9600 ± 0.0182	0.9527 ± 0.0271	0.9685 ± 0.0192	0.9612 ± 0.0124
	Prec	0.9853 ± 0.0111	0.9778 ± 0.0213	0.9791 ± 0.0128	0.9779 ± 0.0201	0.9771 ± 0.0158	0.9607 ± 0.0250
Fusion #3	ACC%	97.46%± 1.24%	96.22%± 1.60%	-	-	-	-
	AUC	0.9925± 0.0093	0.9920 ± 0.0047	-	-	-	-

Continues on the next page

Table B.3: Classification performance on Epistroma dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
	Sen	0.9769 $\pm$ 0.0156	0.9551 $\pm$ 0.0229	-	-	-	-
	Prec	0.9809 $\pm$ 0.0165	0.9814 $\pm$ 0.0103	-	-	-	-

Table B.4: Classification performance on Warwick-QU dataset using shearlet-based descriptors.

		RP + Magnitude		RP		Magnitude	
Method	P	SVM	DTB	SVM	DTB	SVM	DTB
(1): Baseline performance of existing shearlet-based methods							
Vo et al.	ACC%	93.35%±5.23%	86.14%±6.10%	92.13%±8.53%	85.55%±6.83%	73.31%±11.27%	73.97%±12.26%
	AUC	0.9802±0.0351	0.9362 ±0.0347	0.9829 ±0.0345	0.9338 ±0.0437	0.8157 ±0.0998	0.8163 ±0.1047
	Sen	0.8929±0.1349	0.8804 ±0.1107	0.8661 ±0.1661	0.8018 ±0.1201	0.6750 ±0.1976	0.6625 ±0.1872
	Prec	0.9639±0.0583	0.8294 ±0.0786	0.9579 ±0.0690	0.8696 ±0.0766	0.7256 ±0.1569	0.7324 ±0.1391
Meshkini and Ghassemian	ACC%	<i>82.54%±8.52%</i>	<i>87.90%±6.25%</i>	<i>86.14%±7.25%</i>	<i>84.23%±9.54%</i>	81.95%±8.73%	84.38%±8.86%
	AUC	<i>0.9462±0.0326</i>	<i>0.9331±0.0592</i>	<i>0.9293±0.0512</i>	<i>0.8800±0.0676</i>	0.9013 ±0.0455	0.9329 ±0.0546
	Sen	<i>0.7214±0.1475</i>	<i>0.7821±0.1380</i>	<i>0.7589±0.1892</i>	<i>0.7286±0.1828</i>	0.7196 ±0.1409	0.8196 ±0.1896
	Prec	<i>0.8707±0.0937</i>	<i>0.9389±0.0811</i>	<i>0.9324±0.0921</i>	<i>0.8973±0.1164</i>	0.8574 ±0.1125	0.8519 ±0.1161
Zhou et al.	ACC%	<i>63.60%±7.56%</i>	<i>75.85%±8.72%</i>	<i>55.15%±8.52%</i>	<i>67.76%±10.21%</i>	58.79%±6.61%	72.21%±9.57%
	AUC	<i>0.6752±0.1375</i>	<i>0.8544±0.1017</i>	<i>0.5771±0.1091</i>	<i>0.6769±0.1280</i>	0.6006 ±0.1252	0.7699 ±0.1103
	Sen	<i>0.4571±0.1358</i>	<i>0.5964±0.1550</i>	<i>0.4054±0.1296</i>	<i>0.4804±0.1944</i>	0.3375 ±0.1814	0.5821 ±0.1092
	Prec	<i>0.6442±0.1660</i>	<i>0.8349±0.1336</i>	<i>0.5157±0.1430</i>	<i>0.6812±0.1553</i>	0.5764 ±0.1998	0.7665 ±0.1601
Dong et al.	ACC%	<i>91.58%±6.33%</i>	<i>86.07%±11.08%</i>	<i>79.96%±6.46%</i>	<i>83.57%±12.54%</i>	83.13%±5.94%	75.74%±5.06%
Continues on the next page							

Continues on the next page

Table B.4: Classification performance on Warwick-QU dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
	AUC	0.9777 ± 0.0372	0.9345 ± 0.0743	0.9234 ± 0.0547	0.9297 ± 0.0829	0.9342 ± 0.0507	0.8575 ± 0.0826
	Sen	0.9589 ± 0.0663	0.8125 ± 0.1831	0.8089 ± 0.1007	0.7536 ± 0.2074	0.8143 ± 0.1381	0.6625 ± 0.1112
	Prec	0.8844 ± 0.1172	0.8661 ± 0.1076	0.7673 ± 0.0976	0.8843 ± 0.1575	0.8252 ± 0.0942	0.7731 ± 0.0968
	(2): Proposed shearlet-based methods for textured bio-medical image classification						
	ACC%	$95.70\% \pm 7.80\%$	$86.76\% \pm 5.35\%$	$92.61\% \pm 8.25\%$	$84.82\% \pm 8.92\%$	$85.40\% \pm 6.56\%$	$78.20\% \pm 13.74\%$
	AUC	0.9860 ± 0.0353	0.9427 ± 0.0347	0.9709 ± 0.0457	0.9372 ± 0.0451	0.9440 ± 0.0462	0.8992 ± 0.0653
CM	Sen	0.9446 ± 0.0983	0.8679 ± 0.1226	0.9304 ± 0.0736	0.7946 ± 0.1404	0.7946 ± 0.0782	0.7214 ± 0.2090
	Prec	0.9589 ± 0.0945	0.8483 ± 0.0682	0.9181 ± 0.1222	0.8691 ± 0.1277	0.8708 ± 0.1060	0.7756 ± 0.1485
LBP	ACC%	$88.49\% \pm 5.25\%$	$86.69\% \pm 7.24\%$	$84.12\% \pm 7.49\%$	$80.00\% \pm 11.99\%$	$89.71\% \pm 9.45\%$	$82.43\% \pm 8.30\%$
	AUC	0.9592 ± 0.0544	0.9418 ± 0.0416	0.9500 ± 0.0670	0.8944 ± 0.0836	0.9419 ± 0.0538	0.9248 ± 0.0388
	Sen	0.8821 ± 0.1152	0.7982 ± 0.1223	0.8518 ± 0.1021	0.7625 ± 0.1762	0.8946 ± 0.1305	0.7304 ± 0.1685
	Prec	0.8770 ± 0.0955	0.9000 ± 0.0905	0.8292 ± 0.1426	0.7921 ± 0.1434	0.8825 ± 0.1080	0.8649 ± 0.1017
LOSIB	ACC%	$92.57\% \pm 9.24\%$	$87.21\% \pm 5.39\%$	$89.52\% \pm 7.85\%$	$87.43\% \pm 10.53\%$	$79.34\% \pm 5.33\%$	$76.80\% \pm 10.66\%$
	AUC	0.9798 ± 0.0395	0.9346 ± 0.0432	0.9671 ± 0.0554	0.9344 ± 0.0570	0.9153 ± 0.0601	0.8679 ± 0.0907
	Sen	0.9446 ± 0.0983	0.7964 ± 0.1155	0.8464 ± 0.1258	0.8411 ± 0.1442	0.7732 ± 0.1465	0.7125 ± 0.1685
	Prec	0.9042 ± 0.1255	0.9177 ± 0.0902	0.9264 ± 0.1119	0.8792 ± 0.1265	0.7983 ± 0.1328	0.7579 ± 0.1123
SFTA	ACC%	$94.52\% \pm 7.47\%$	$87.76\% \pm 5.95\%$	$92.68\% \pm 7.06\%$	$84.23\% \pm 10.17\%$	$85.99\% \pm 7.09\%$	$82.39\% \pm 6.18\%$
	AUC	0.9846 ± 0.0350	0.9353 ± 0.0469	0.9739 ± 0.0416	0.9192 ± 0.0937	0.9515 ± 0.0410	0.9119 ± 0.0692
	Sen	0.9196 ± 0.0968	0.8196 ± 0.1576	0.9179 ± 0.0710	0.7821 ± 0.1301	0.7946 ± 0.0979	0.8286 ± 0.1203

Continues on the next page

Table B.4: Classification performance on Warwick-QU dataset using shearlet-based descriptors.

Method	P	RP + Magnitude		RP		Magnitude	
		SVM	DTB	SVM	DTB	SVM	DTB
	Prec	0.9589 $\pm$ 0.0945	0.9107 $\pm$ 0.0869	0.9246 $\pm$ 0.1059	0.8767 $\pm$ 0.1447	0.8798 $\pm$ 0.0881	0.8186 $\pm$ 0.1372
(3): Integrating shearlet-based existing techniques with our proposed methods							
Fusion #1	ACC%	95.11% $\pm$ 7.67%	88.42% $\pm$ 7.84%	92.65% $\pm$ 8.26%	92.06% $\pm$ 7.23%	87.83% $\pm$ 5.07%	79.45% $\pm$ 10.88%
	AUC	0.9860 $\pm$ 0.0353	0.9527 $\pm$ 0.0475	0.9721 $\pm$ 0.0511	0.9645 $\pm$ 0.0603	0.9651 $\pm$ 0.0302	0.8867 $\pm$ 0.0704
	Sen	0.9321 $\pm$ 0.0985	0.8750 $\pm$ 0.0807	0.9321 $\pm$ 0.0718	0.9196 $\pm$ 0.0968	0.8661 $\pm$ 0.0597	0.7429 $\pm$ 0.1543
	Prec	0.9589 $\pm$ 0.0945	0.8833 $\pm$ 0.1277	0.9181 $\pm$ 0.1222	0.9125 $\pm$ 0.1006	0.8724 $\pm$ 0.0991	0.8111 $\pm$ 0.1621
Fusion #2	ACC%	96.29%$\pm$ 7.89%	90.99% $\pm$ 7.13%	92.02% $\pm$ 7.86%	90.81% $\pm$ 8.42%	87.79% $\pm$ 5.90%	79.26% $\pm$ 8.62%
	AUC	0.9860$\pm$ 0.0353	0.9580 $\pm$ 0.0446	0.9734 $\pm$ 0.0525	0.9696 $\pm$ 0.0369	0.9601 $\pm$ 0.0322	0.8843 $\pm$ 0.0919
	Sen	0.9571$\pm$ 0.0964	0.9232 $\pm$ 0.0887	0.9304 $\pm$ 0.0736	0.9446 $\pm$ 0.0983	0.8786 $\pm$ 0.0436	0.7411 $\pm$ 0.1399
	Prec	0.9589$\pm$ 0.0945	0.8913 $\pm$ 0.1026	0.9069 $\pm$ 0.1189	0.8783 $\pm$ 0.1274	0.8682 $\pm$ 0.1247	0.7994 $\pm$ 0.1370
Fusion #3	ACC%	94.49% $\pm$ 7.47%	85.48% $\pm$ 9.86%	-	-	-	-
	AUC	0.9860 $\pm$ 0.0353	0.9523 $\pm$ 0.0570	-	-	-	-
	Sen	0.9304 $\pm$ 0.0998	0.8161 $\pm$ 0.1473	-	-	-	-
	Prec	0.9478 $\pm$ 0.0957	0.8718 $\pm$ 0.1454	-	-	-	-

References

- [1] Anees Abrol, Manish Bhattarai, Alex Fedorov, Yuhui Du, Sergey Plis, Vince Calhoun, Alzheimer’s Disease Neuroimaging Initiative, et al. Deep residual learning for neuroimaging: An application to predict progression to alzheimer’s disease. *Journal of Neuroscience Methods*, page 108701, 2020.
- [2] Euijoon Ahn, Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. *IEEE Transactions on Medical Imaging*, 39(7):2385–2394, 2020.
- [3] Shaimaa Al-Janabi, André Huisman, and Paul J Van Diest. Digital pathology: current status and future perspectives. *Histopathology*, 61(1):1–9, 2012.
- [4] Taha J Alhindi, Shivam Kalra, Ka Hin Ng, Anika Afrin, and Hamid R Tizhoosh. Comparing LBP, HOG and deep features for classification of histopathology images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.
- [5] Sadiq Alinsaif and Jochen Lang. Shearlet-based techniques for histological image classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1424–1431. IEEE, 2019.
- [6] Sadiq Alinsaif and Jochen Lang. Histological image classification using deep features and transfer learning. In *2020 17th Conference on Computer and Robot Vision (CRV)*, pages 101–108. IEEE, 2020.
- [7] Sadiq Alinsaif and Jochen Lang. Texture features in the shearlet domain for histopathological image classification. *BMC Medical Informatics and Decision Making*, 20(14):1–19, 2020.
- [8] Sadiq Alinsaif and Jochen Lang. 3D shearlet-based descriptors combined with deep features for the classification of Alzheimer’s disease based on MRI data. 2021. Under review.
- [9] Sadiq Alinsaif and Jochen Lang. Histopathological image classification: a review of datasets and techniques. 2021. Under review.

- [10] Md Zahangir Alom, Theus Aspiras, Tarek M Taha, and Vijayan K Asari. Histopathological image classification with deep convolutional neural networks. In *Applications of Machine Learning*, volume 11139, page 111390X. International Society for Optics and Photonics, 2019.
- [11] Md Zahangir Alom, Chris Yakopcic, Mst Shamima Nasrin, Tarek M Taha, and Vijayan K Asari. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *Journal of Digital Imaging*, 32(4):605–617, 2019.
- [12] Tooba Altaf, Syed Muhammad Anwar, Nadia Gul, Muhammad Nadeem Majeed, and Muhammad Majid. Multi-class Alzheimer’s disease classification using image and clinical features. *Biomedical Signal Processing and Control*, 43:64–74, 2018.
- [13] Jacopo Annese. The importance of combining mri and large-scale digital histology in neuroimaging studies of brain connectivity and disease. *Frontiers in Neuroinformatics*, 6:13, 2012.
- [14] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.
- [15] Ruqayya Awan, Navid Alemi Koohbanani, Muhammad Shaban, Anna Lisowska, and Nasir Rajpoot. Context-aware learning using transferable features for classification of breast cancer histology images. In *International Conference Image Analysis and Recognition*, pages 788–795. Springer, 2018.
- [16] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.
- [17] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.
- [18] Raquel Bello-Cerezo, Francesco Bianconi, Francesco Di Maria, Paolo Napoletano, and Fabrizio Smeraldi. Comparative evaluation of hand-crafted image descriptors vs. off-the-shelf CNN-based features for colour texture classification under ideal and realistic conditions. *Applied Sciences*, 9(4):738, 2019.
- [19] Aïcha BenTaieb and Ghassan Hamarneh. Uncertainty driven multi-loss fully convolutional networks for histopathology. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 155–163. Springer, 2017.

- [20] Abi Berger. How does it work?: Magnetic resonance imaging. *BMJ: British Medical Journal*, 324(7328):35, 2002.
- [21] Francesco Bianconi, Alberto Álvarez-Larrán, and Antonio Fernández. Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing*, 154:119–126, 2015.
- [22] Francesco Bianconi, Raquel Bello-Cerezo, and Paolo Napoletano. Improved opponent color local binary patterns: an effective local image descriptor for color texture classification. *Journal of Electronic Imaging*, 27(1):011002, 2017.
- [23] Francesco Bianconi and Antonio Fernández. An appendix to “texture databases—a comprehensive survey”. *Pattern Recognition Letters*, 45:33–38, 2014.
- [24] Francesco Bianconi, Antonio Fernández, Elena González, and Stefano A Saetta. Performance analysis of colour descriptors for parquet sorting. *Expert Systems with Applications*, 40(5):1636–1644, 2013.
- [25] Nadia Brancati, Maria Frucci, and Daniel Riccio. Multi-classification of breast cancer histology images by using a fine-tuning strategy. In *International Conference Image Analysis and Recognition*, pages 771–778. Springer, 2018.
- [26] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [28] Daniel O Tambasco Bruno, Marcelo Z do Nascimento, Rodrigo P Ramos, Valério R Batista, Leandro A Neves, and Alessandro S Martins. LBP operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues. *Expert Systems with Applications*, 55:329–340, 2016.
- [29] Lingyun Cai, Xin Wang, Yuanyuan Wang, Yi Guo, Jinhua Yu, and Yi Wang. Robust phase-based texture descriptor for classification of breast ultrasound images. *Biomedical Engineering Online*, 14(1):26, 2015.
- [30] Emmanuel J Candès and David L Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(2):219–266, 2004.
- [31] Ping Cao, Jie Gao, and Zuping Zhang. Multi-view based multi-model learning for MCI diagnosis. *Brain Sciences*, 10(3):181, 2020.
- [32] Silvia Cascianelli, Raquel Bello-Cerezo, Francesco Bianconi, Mario L Fravolini, Mehdi Belal, Barbara Palumbo, and Jakob N Kather. Dimensionality reduction strategies for CNN-based classification of histopathological images. In *International Conference on Intelligent Interactive Multimedia Systems and Services*, pages 21–30. Springer, 2018.

- [33] Subhankar Chatteraj, Souvik Pratiher, Rajdeep Mukherjee, Saikat Ghosh, Arnab Chakraborty, Diptiman Hazra, and Sawon Pratiher. Efficacy of deep convolutional neural network features on histological manifold for robust breast carcinoma detection. In *Optics in Health Care and Biomedical Optics VIII*, volume 10820, page 108203Q. International Society for Optics and Photonics, 2018.
- [34] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [35] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
- [36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- [37] Sai Saketh Chennamsetty, Mohammed Safwan, and Varghese Alex. Classification of breast cancer histology image using ensemble of pre-trained neural networks. In *International Conference Image Analysis and Recognition*, pages 804–811. Springer, 2018.
- [38] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.
- [39] Garry Choy, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E Samir, Oleg S Pianykh, J Raymond Geis, Pari V Pandharipande, James A Brink, and Keith J Dreyer. Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318–328, 2018.
- [40] David A Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1):45–62, 2002.
- [41] Richard W Connors and Charles A Harlow. A theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3):204–222, 1980.
- [42] Alceu Ferraz Costa, Gabriel Humpire-Mamani, and Agma Juci Machado Traina. An efficient algorithm for fractal analysis of textures. In *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on*, pages 39–46. IEEE, 2012.
- [43] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
- [44] Angel Cruz-Roa, Juan C Caicedo, and Fabio A González. Visual pattern mining in histology image collections using bag of features. *Artificial Intelligence in Medicine*, 52(2):91–106, 2011.

- [45] Claudio Cusano, Paolo Napoletano, and Raimondo Schettini. Combining local binary patterns and local color contrast for texture classification under varying illumination. *JOSA A*, 31(7):1453–1461, 2014.
- [46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [47] Cecilia Di Ruberto, Giuseppe Fodde, and Lorenzo Putzu. Comparison of statistical features for medical colour image classification. In Lazaros Nalpantidis, Volker Krüger, Jan-Olof Eklundh, and Antonios Gasteratos, editors, *Computer Vision Systems*, pages 3–13, Cham, 2015. Springer International Publishing.
- [48] Minh N Do and Martin Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.
- [49] Marcelo Zanchetta do Nascimento, Alessandro Santana Martins, Thaína Aparecida Azevedo Tosta, and Leandro Alves Neves. Lymphoma images analysis using morphological and non-morphological descriptors for classification. *Computer Methods and Programs in Biomedicine*, 163:65–77, 2018.
- [50] Yongsheng Dong, Dacheng Tao, Xuelong Li, Jinwen Ma, and Jiexin Pu. Texture classification and retrieval using shearlets and linear regression. *IEEE Transactions on Cybernetics*, 45(3):358–369, 2015.
- [51] Florentino Luciano Caetano dos Santos, Michelangelo Paci, Loris Nanni, Sheryl Brahnam, and Jari Hyttinen. Computer vision for virus image classification. *Biosystems Engineering*, 138:11–22, 2015.
- [52] Luiz Fernando Segato dos Santos, Leandro Alves Neves, Guilherme Botazzo Rozendo, Matheus Gonçalves Ribeiro, Marcelo Zanchetta do Nascimento, and Thaína Aparecida Azevedo Tosta. Multidimensional and fuzzy sample entropy (sampenmf) for quantifying H&E histological images of colorectal cancer. *Computers in Biology and Medicine*, 103:148–160, 2018.
- [53] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.
- [54] Dennis Dunn and William E Higgins. Optimal Gabor filters for texture segmentation. *IEEE Transactions on Image Processing*, 4(7):947–964, 1995.
- [55] Glenn Easley, Demetrio Labate, and Wang-Q Lim. Sparse directional image representations using the discrete shearlet transform. *Applied and Computational Harmonic Analysis*, 25(1):25–46, 2008.

- [56] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. Deep learning to detect Alzheimer’s disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 187:105242, 2020.
- [57] Livia Faes, Xiaoxuan Liu, Aditya Kale, Alice Bruynseels, Mohith Shamdas, Gabriella Moraes, Dun Jack Fu, Siegfried K Wagner, Christoph Kern, Joseph RE Ledsam, et al. Deep learning under scrutiny: performance against health care professionals in detecting diseases from medical imaging-systematic review and meta-analysis. 2019.
- [58] Farnaz Farokhian, Iman Beheshti, Daichi Sone, and Hiroshi Matsuda. Comparing CAT12 and VBM8 for detecting brain morphological abnormalities in temporal lobe epilepsy. *Frontiers in Neurology*, 8:428, 2017.
- [59] Moh Faturrahman, Ito Wasito, Nurul Hanifah, and Ratna Mufidah. Structural MRI classification for Alzheimer’s disease detection using deep belief network. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, pages 37–42. IEEE, 2017.
- [60] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- [61] Jinwang Feng, Shao-Wu Zhang, Luonan Chen, Alzheimer’s Disease Neuroimaging Initiative (ADNI, et al. Identification of Alzheimer’s disease based on wavelet transformation energy feature of the structural MRI image and NN classifier. *Artificial Intelligence in Medicine*, 108:101940, 2020.
- [62] Daniel C Fernandez, Rohit Bhargava, Stephen M Hewitt, and Ira W Levin. Infrared spectroscopic imaging for histopathologic recognition. *Nature Biotechnology*, 23(4):469, 2005.
- [63] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.
- [64] Irene Fondón, Auxiliadora Sarmiento, Ana Isabel García, María Silvestre, Catarina Eloy, António Polónia, and Paulo Aguiar. Automatic classification of tissue malignancy for breast carcinoma diagnosis. *Computers in Biology and Medicine*, 96:41–51, 2018.
- [65] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufmann.
- [66] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001.

- [67] Oscar García-Olalla, Enrique Alegre, Laura Fernández-Robles, and Víctor González-Castro. Local oriented statistics information booster (LOSIB) for texture classification. In *2014 22nd International Conference on Pattern Recognition*, pages 1114–1119. IEEE, 2014.
- [68] Christian Gaser and Florian Kurth. Manual computational anatomy toolbox-cat12. *Structural Brain Mapping Group at the Departments of Psychiatry and Neurology, University of Jena*, 2017.
- [69] Elisa Drelie Gelasca, Boguslaw Obara, Dmitry Fedorov, Kristian Kvilekval, and BS Manjunath. A biosegmentation benchmark for evaluation of bioimage analysis methods. *BMC Bioinformatics*, 10(1):368, 2009.
- [70] Aditya Golatkar, Deepak Anand, and Amit Sethi. Classification of breast cancer histology using deep learning. In *International Conference Image Analysis and Recognition*, pages 837–844. Springer, 2018.
- [71] Germán González, Rodrigo Nava, and Boris Escalante-Ramírez. A comparative study on discrete shmaliiy moments and their texture-based applications. *Mathematical Problems in Engineering*, 2018, 2018.
- [72] Stefano Graña, Loris Nanni, Sheryl Brahnam, and Emanuele Menegatti. Texture descriptors based on dijkstra’s algorithm for medical image analysis. *Innov. Med. Healthcare*, 207:74, 2015.
- [73] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI’11, pages 266–273. AUAI Press, 2011.
- [74] David Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), and Web*, 2, 2017.
- [75] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.
- [76] Vibha Gupta and Arnav Bhavsar. Sequential Modeling of Deep Features for Breast Cancer Histopathological Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2254–2261, 2018.
- [77] Robert M Haralick, Karthikeyan Shanmugam, and Its’ Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973.
- [78] Harry H Harman. *Modern factor analysis*. University of Chicago press, 1976.
- [79] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In *Advances in neural information processing systems*, pages 507–513, 1998.

- [80] Simon Haykin. A comprehensive foundation. *Neural Networks*, 2(2004):41, 2004.
- [81] Jiangping He, Hongwei Ji, and Xin Yang. Rotation invariant texture descriptor using local shearlet-based energy histograms. *IEEE Signal Processing Letters*, 20(9):905–908, 2013.
- [82] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [83] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [84] Tim D Hewitson and Ian A Darby. *Histology protocols*. Springer, 2010.
- [85] Marcia Hon and Naimul Mefraz Khan. Towards alzheimer’s disease classification through transfer learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1166–1169. IEEE, 2017.
- [86] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [87] Shahera Hossain and Seiichi Serikawa. Texture databases—a comprehensive survey. *Pattern Recognition Letters*, 34(15):2007–2022, 2013.
- [88] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.
- [89] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [90] Yue Huang, Han Zheng, Chi Liu, Xinghao Ding, and Gustavo K Rohde. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1625–1632, 2017.
- [91] Zhongling Huang, Zongxu Pan, and Bin Lei. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sensing*, 9(9):907, 2017.
- [92] Anne Humeau-Heurtier. Texture feature extraction methods: A survey. *IEEE Access*, 7:8975–9000, 2019.
- [93] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

- [94] Tomas Iesmantas and Robertas Alzbutas. Convolutional capsule network for classification of breast cancer histology images. In *International Conference Image Analysis and Recognition*, pages 853–860. Springer, 2018.
- [95] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE Reviews in Biomedical Engineering*, 7:97–114, 2014.
- [96] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.
- [97] Rachna Jain, Nikita Jain, Akshay Aggarwal, and D Jude Hemanth. Convolutional neural network based Alzheimer’s disease classification from magnetic resonance brain images. *Cognitive Systems Research*, 57:147–159, 2019.
- [98] Debesh Jha, Ji-In Kim, and Goo-Rak Kwon. Diagnosis of Alzheimer’s disease using dual-tree complex wavelet transform, PCA, and feed-forward neural network. *Journal of Healthcare Engineering*, 2017, 2017.
- [99] Yun Jiang, Li Chen, Hai Zhang, and Xiao Xiao. Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *PLoS ONE*, 14(3), 2019.
- [100] Ian Jolliffe. *Principal component analysis*. Springer, 2011.
- [101] Padmaja Jonnalagedda, Daniel Schmolze, and Bir Bhanu. MVPNets: Multi-viewing path deep learning neural networks for magnification invariant diagnosis in breast cancer. In *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 189–194. IEEE, 2018.
- [102] Juho Kannala and Esa Rahtu. BSIF: Binarized statistical image features. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1363–1366. IEEE, 2012.
- [103] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, 2016.
- [104] Asifullah Khan, Anabia Sohail, Umme Zahoor, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, pages 1–62, 2020.
- [105] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using

- pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.
- [106] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
 - [107] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019.
 - [108] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.
 - [109] Ismaël Koné and Lahsen Boulmane. Hierarchical ResNeXt models for breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, pages 796–803. Springer, 2018.
 - [110] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, 1997.
 - [111] K Krishna and Narasimha M Murty. Genetic k-means algorithm. *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, 29(3):433–439, 1999.
 - [112] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
 - [113] Abhinav Kumar, Sanjay Kumar Singh, Sonal Saxena, K Lakshmanan, Arun Kumar Sangaiah, Himanshu Chauhan, Sameer Shrivastava, and Raj Kumar Singh. Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Information Sciences*, 508:405–421, 2020.
 - [114] Meghana Dinesh Kumar, Morteza Babaie, Shujin Zhu, Shivam Kalra, and Hamid R Tizhoosh. A comparative study of CNN, BoVW and LBP for classification of histopathological images. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2017.
 - [115] Gitta Kutyniok and Demetrio Labate. Introduction to shearlets. *Shearlets*, pages 1–38, 2012.
 - [116] Gitta Kutyniok, Wang-Q Lim, and Gabriele Steidl. Shearlets: theory and applications. *GAMM-Mitteilungen*, 37(2):259–280, 2014.
 - [117] Gitta Kutyniok, Morteza Shahram, and Xiaosheng Zhuang. Shearlab: A rational design of a digital parabolic scaling algorithm. *SIAM Journal on Imaging Sciences*, 5(4):1291–1332, 2012.

- [118] Ko-Hua Lai and Meng-Chou Chang. Classification of tissue types in histological colorectal cancer images using residual networks. In *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pages 299–300. IEEE, 2019.
- [119] ZhiFei Lai and HuiFang Deng. Medical images classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron. *Computational intelligence and neuroscience*, 2018.
- [120] Hung Le Minh, Manh Mai Van, and Tran Van Lang. Deep feature fusion for breast cancer diagnosis on histopathology images. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. IEEE, 2019.
- [121] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [122] Subin Lee, Hyunna Lee, and Ki Woong Kim. Magnetic resonance imaging texture predicts progression to dementia due to Alzheimer disease earlier than hippocampal volume. *Journal of Psychiatry & Neuroscience: JPN*, 45(1):7, 2020.
- [123] X. Li and K. N. Plataniotis. A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. *IEEE Transactions on Biomedical Engineering*, 62(7):1862–1873, July 2015.
- [124] Xingyu Li and Konstantinos N Plataniotis. Color model comparative analysis for breast cancer diagnosis using h and e stained images. In *Medical Imaging 2015: Digital Pathology*, volume 9420, page 94200L. International Society for Optics and Photonics, 2015.
- [125] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer’s disease diagnosis using structural mri. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [126] Nina Linder, Juho Konsti, Riku Turkki, Esa Rahtu, Mikael Lundin, Stig Nordling, Caj Haglund, Timo Ahonen, Matti Pietikäinen, and Johan Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology*, 7(1):22, 2012.
- [127] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. SIFT flow: Dense correspondence across different scenes. In *European Conference on Computer Vision*, pages 28–42. Springer, 2008.
- [128] Cheng-Lin Liu and Hiroshi Sako. Class-specific feature polynomial classifier for pattern classification and its application to handwritten numeral recognition. *Pattern Recognition*, 39(4):669–681, 2006.
- [129] Dong Liu, Yaohui Liu, Shanglin Li, Weiqing Li, and Luda Wang. Fusion of hand-crafted and deep features for medical image classification. In *Journal of Physics: Conference Series*, volume 1345, page 022052. IOP Publishing, 2019.

- [130] Xiaoming Liu, Jun Liu, Zhilin Feng, Xin Xu, and J Tang. Mass classification in mammogram with semi-supervised relief based feature selection. In *Fifth International Conference on Graphic and Image Processing (ICGIP 2013)*, volume 9069, page 90691B. International Society for Optics and Photonics, 2014.
- [131] Zhiqiang Liu, Paul Chow, Jinwei Xu, Jingfei Jiang, Yong Dou, and Jie Zhou. A uniform architecture design for accelerating 2d and 3d cnns on fpgas. *Electronics*, 8(1):65, 2019.
- [132] Wei-Yin Loh and Yu-Shan Shih. Split selection methods for classification trees. *Statistica sinica*, pages 815–840, 1997.
- [133] David G Lowe et al. Object recognition from local scale-invariant features. In *ICCV*, volume 99, pages 1150–1157, 1999.
- [134] Alessandra Lumini, Loris Nanni, and Sheryl Brahnam. Multilayer descriptors for medical image classification. *Computers in Biology and Medicine*, 72:239–247, 2016.
- [135] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.
- [136] Gianluca Maguolo, Loris Nanni, and Stefano Ghidoni. Ensemble of convolutional neural networks trained with different activation functions. *arXiv preprint arXiv:1905.02473*, 2019.
- [137] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.
- [138] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (PPMI). *Progress in Neurobiology*, 95(4):629–635, 2011.
- [139] Claudia Mazo, Jose Bernal, Maria Trujillo, and Enrique Alegre. Transfer learning for classification of cardiovascular tissues in histological images. *Computer Methods and Programs in Biomedicine*, 165:69–76, 2018.
- [140] Alfonso Medela and Artzai Picon. Constellation loss: Improving the efficiency of deep metric learning loss functions for optimal embedding. *arXiv preprint arXiv:1905.10675*, 2019.
- [141] Ricardo Mendoza-Léon, John Puentes, Luis Felipe Uriza, and Marcela Hernández Hoyos. Single-slice Alzheimer’s disease classification and disease regional analysis with supervised switching autoencoders. *Computers in Biology and Medicine*, 116:103527, 2020.

- [142] Khatere Meshkini and Hassan Ghassemian. Texture classification using Shearlet transform and GLCM. In *Electrical Engineering (ICEE), 2017 Iranian Conference on*, pages 1845–1850. IEEE, 2017.
- [143] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.
- [144] B. S. Shajee Mohan and C. Chandra Sekhar. Class-specific mahalanobis distance metric learning for biological image classification. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, pages 240–248, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- [145] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [146] Abdullah-Al Nahid and Yinan Kong. Histopathological breast-image classification using local and frequency domains by convolutional neural network. *Information*, 9(1):19, 2018.
- [147] Loris Nanni, Sheryl Brahnam, Christian Salvatore, Isabella Castiglioni, Alzheimer’s Disease Neuroimaging Initiative, et al. Texture descriptors and voxels for the early diagnosis of Alzheimer’s disease. *Artificial Intelligence in Medicine*, 97:19–26, 2019.
- [148] Loris Nanni and Stefano Ghidoni. How could a subcellular image, or a painting by van gogh, be similar to a great white shark or to a pizza? *Pattern Recognition Letters*, 85:1–7, 2017.
- [149] Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172, 2017.
- [150] Loris Nanni, Michelangelo Paci, Sheryl Brahnam, and Stefano Ghidoni. An ensemble of visual features for gaussians of local descriptors and non-binary coding for texture descriptors. *Expert Systems with Applications*, 82:27–39, 2017.
- [151] Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS ONE*, 11(2):e0149399, 2016.
- [152] Loris Nannia, Stefano Ghidoni, and Sheryl Brahnam. Ensemble of convolutional neural networks for bioimage classification. *Applied Computing and Informatics*, 2020.
- [153] R. Nava, G. González, J. Kybic, and B. Escalante-Ramírez. Characterization of hematologic malignancies based on discrete orthogonal moments. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Dec 2016.

- [154] Rodrigo Nava, Germán González, Jan Kybic, and Boris Escalante-Ramírez. Classification of tumor epithelium and stroma in colorectal cancer based on discrete Tchebichef moments. In *Workshop on Clinical Image-Based Procedures*, pages 79–87. Springer, 2015.
- [155] Kamyar Nazari, Azad Aminpour, and Mehran Ebrahimi. Two-stage convolutional neural network for breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, pages 717–726. Springer, 2018.
- [156] Ryusuke Nosaka, Chendra Hadi Suryanto, and Kazuhiro Fukui. Rotation invariant co-occurrence among adjacent LBPs. In *Asian Conference on Computer Vision*, pages 15–25. Springer, 2012.
- [157] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987, 2002.
- [158] Ville Ojansivu and Janne Heikkilä. Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer, 2008.
- [159] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [160] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.
- [161] Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe, and Ilya G Goldberg. Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):1003–1013, 2010.
- [162] Dan Pan, An Zeng, Longfei Jia, Yin Huang, Tory Frizzell, and Xiaowei Song. Early detection of Alzheimer’s disease using magnetic resonance imaging: A novel approach combining convolutional neural networks and ensemble learning. *Frontiers in Neuroscience*, 14, 2020.
- [163] Yue Pan, Li Liu, Longfei Yang, and Yizheng Wang. Texture feature extracting method based on local relative phase binary pattern. In *2016 5th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 749–753. IEEE, 2016.
- [164] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991.
- [165] Gil Patrus Pena and Joséde Souza Andrade-Filho. How does a pathologist make a diagnosis? *Archives of Pathology & Laboratory Medicine*, 133(1):124–132, 2009.

- [166] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer, 2010.
- [167] Artem Pimkin, Gleb Makarchuk, Vladimir Kondratenko, Maxim Pisov, Egor Krivov, and Mikhail Belyaev. Ensembling neural networks for digital pathology images classification and segmentation. In *International Conference Image Analysis and Recognition*, pages 877–886. Springer, 2018.
- [168] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.
- [169] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
- [170] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [171] Jia Qu, Nobuyuki Hiruta, Kensuke Terai, Hirokazu Nosato, Masahiro Murakawa, and Hidenori Sakanashi. Gastric Pathology Image Classification Using Stepwise Fine-Tuning for Deep Neural Networks. *Journal of Healthcare Engineering*, 2018, 2018.
- [172] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
- [173] Venkatesan Rajinikanth, Alex Noel Joseph Raj, Krishnan Palani Thanaraj, and Ganesh R Naik. A customized vgg19 network with concatenation of deep and hand-crafted features for brain tumor detection. *Applied Sciences*, 10(10):3429, 2020.
- [174] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition*, pages 737–744. Springer, 2018.
- [175] Geraldo L Bezerra Ramalho, Daniel S Ferreira, Pedro P Rebouças Filho, and Fátima N Sombra de Medeiros. Rotation-invariant feature extraction using a structural co-occurrence matrix. *Measurement*, 94:406–415, 2016.
- [176] Rafael Reisenhofer. *The complex shearlet transform and applications to image quality assessment*. PhD thesis, Master’s thesis, Technische Universität Berlin, 2014.
- [177] Matheus Gonçalves Ribeiro, Leandro Alves Neves, Marcelo Zanchetta do Nascimento, Guilherme Freire Roberto, Alessandro Santana Martins, and Thaína Aparecida Azevedo Tosta. Classification of colorectal cancer based on the association of multidimensional and multiresolution features. *Expert Systems with Applications*, 120:262–278, 2019.

- [178] Guilherme F Roberto, Leandro A Neves, Marcelo Z Nascimento, Thaína AA Tosta, Leonardo C Longo, Alessandro S Martins, and Paulo R Faria. Features based on the percolation theory for quantification of non-hodgkin lymphomas. *Computers in Biology and Medicine*, 91:135–147, 2017.
- [179] Giorgio Roffo, Simone Melzi, Umberto Castellani, and Alessandro Vinciarelli. Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1398–1406, 2017.
- [180] Lucy Balian Rorke. Pathologic diagnosis as the gold standard. *Cancer*, 79(4):665–667, 1997.
- [181] Kaushiki Roy, Debapriya Banik, Debotosh Bhattacharjee, and Mita Nasipuri. Patch-based system for classification of breast histology images using deep learning. *Computerized Medical Imaging and Graphics*, 71:90–103, 2019.
- [182] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011.
- [183] Sebastian Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.
- [184] Christian Salvatore, Antonio Cerasa, Petronilla Battista, Maria Carla Gilardi, Aldo Quattrone, and Isabella Castiglioni. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer’s disease: a machine learning approach. *Frontiers in Neuroscience*, 9:307, 2015.
- [185] Christian Salvatore, Antonio Cerasa, and Isabella Castiglioni. MRI characterizes the progressive course of AD and predicts conversion to Alzheimer’s dementia 24 months before probable diagnosis. *Frontiers in Aging Neuroscience*, 10:135, 2018.
- [186] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [187] Auxiliadora Sarmiento and Irene Fondón. Automatic breast cancer grading of histological images based on colour and texture descriptors. In *International Conference Image Analysis and Recognition*, pages 887–894. Springer, 2018.
- [188] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.
- [189] Hannes Schulz and Sven Behnke. Deep learning. *KI-Künstliche Intelligenz*, 26(4):357–363, 2012.

- [190] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm—an open source utility for biological image analysis. *Source code for Biology and Medicine*, 3(1):13, 2008.
- [191] Lior Shamir, Nikita Orlov, David Mark Eckley, Tomasz J Macura, and Ilya G Goldberg. IICBU 2008: a proposed benchmark suite for biological image analysis. *Medical & Biological Engineering & Computing*, 46(9):943–947, 2008.
- [192] Shallu Sharma and Rajesh Mehra. Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express*, 4(4):247–254, 2018.
- [193] Shallu Sharma and Rajesh Mehra. Effect of layer-wise fine-tuning in magnification-dependent classification of breast cancer histopathological image. *The Visual Computer*, pages 1–15, 2019.
- [194] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.
- [195] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [196] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502, 2017.
- [197] Fabrizio Smeraldi, Francesco Bianconi, Antonio Fernández, and Elena González. Partial order rank features in colour space. *Applied Sciences*, 10(2):499, 2020.
- [198] L-K Soh and Costas Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):780–795, 1999.
- [199] Tiecheng Song, Hongliang Li, Fanman Meng, Qingbo Wu, and Jianfei Cai. LETRIST: Locally encoded transform feature histogram for rotation-invariant texture classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(7):1565–1579, 2018.
- [200] Y. Song, Q. Li, H. Huang, D. Feng, M. Chen, and W. Cai. Low dimensional representation of fisher vectors for microscopy image classification. *IEEE Transactions on Medical Imaging*, 36(8):1636–1649, Aug 2017.
- [201] Yang Song, Tom Weidong Cai, Heng Huang, David Dagan Feng, Yue Wang, and Mei Chen. Bioimage classification with subcategory discriminant transform of high dimensional visual descriptors. In *BMC Bioinformatics*, 2016.
- [202] Yang Song, Hang Chang, Yang Gao, Sidong Liu, Donghao Zhang, Junen Yao, Wojciech Chrzanowski, and Weidong Cai. Feature learning with component selective encoding for histopathology image classification. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 257–260. IEEE, 2018.

- [203] James Sorace, Denise R Aberle, Dena Elimam, Silvana Lawvere, Ossama Tawfik, and W Dean Wallace. Integrating pathology and radiology disciplines: an emerging opportunity? *BMC Medicine*, 10(1):100, 2012.
- [204] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.
- [205] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2560–2567. IEEE, 2016.
- [206] Mikkel B Stegmann, Karl Sjöstrand, and Rasmus Larsen. Sparse modeling of landmark and texture variability using the orthomax criterion. In *Medical Imaging 2006: Image Processing*, volume 6144, page 61441G. International Society for Optics and Photonics, 2006.
- [207] Jyothi Subramanian and Richard Simon. Overfitting in prediction models—is it a problem only in high dimensions? *Contemporary Clinical Trials*, 36(2):636–641, 2013.
- [208] KN Sukhia, MM Riaz, A Ghafoor, SS Ali, and N Iltaf. Content-based histopathological image retrieval using multi-scale and multichannel decoder based LTP. *Biomedical Signal Processing and Control*, 54:101582, 2019.
- [209] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [210] Daniela F Taino, Matheus G Ribeiro, Guilherme Freire Roberto, Geraldo FD Zafalon, Marcelo Zanchetta do Nascimento, Thaína A Tosta, Alessandro S Martins, and Leandro A Neves. A model based on genetic algorithm for colorectal cancer diagnosis. In *Iberoamerican Congress on Pattern Recognition*, pages 504–513. Springer, 2019.
- [211] Xiaoyang Tan and William Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, 2010.
- [212] Xiaoou Tang. Texture information in run-length matrices. *IEEE Transactions on Image Processing*, 7(11):1602–1609, 1998.
- [213] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.
- [214] M Tanveer, B Richhariya, RU Khan, AH Rashid, P Khanna, M Prasad, and CT Lin. Machine learning techniques for the diagnosis of Alzheimer’s disease: A review.

ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s):1–35, 2020.

- [215] Phan Thanh Noi and Martin Kappas. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors*, 18(1):18, 2018.
- [216] Ryan Therrien and Scott Doyle. Role of training data variability on classifier performance and generalizability. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058109. International Society for Optics and Photonics, 2018.
- [217] Yeeleng S Vang, Zhen Chen, and Xiaohui Xie. Deep learning framework for multi-class breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, pages 914–922. Springer, 2018.
- [218] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.
- [219] Sulaiman Vesal, Nishant Ravikumar, AmirAbbas Davari, Stephan Ellmann, and Andreas Maier. Classification of breast cancer histology images using transfer learning. In *International Conference Image Analysis and Recognition*, pages 812–819. Springer, 2018.
- [220] An PN Vo, Soontorn Oraintara, and Truong T Nguyen. Using phase and magnitude information of the complex directional filter bank for texture image retrieval. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV–61. IEEE, 2007.
- [221] Duc My Vo, Ngoc-Quang Nguyen, and Sang-Woong Lee. Classification of breast cancer histology images using incremental boosting convolution networks. *Information Sciences*, 482:123–138, 2019.
- [222] Chaofeng Wang, Jun Shi, Qi Zhang, and Shihui Ying. Histopathological image classification with bilinear convolutional neural networks. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 4050–4053. IEEE, 2017.
- [223] Shui-Hua Wang, Preetha Phillips, Yuxiu Sui, Bin Liu, Ming Yang, and Hong Cheng. Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *Journal of Medical Systems*, 42(5):85, 2018.
- [224] Yaqi Wang, Lingling Sun, Kaiqiang Ma, and Jiannan Fang. Breast cancer microscope image classification based on CNN with image deformation. In *International Conference Image Analysis and Recognition*, pages 845–852. Springer, 2018.
- [225] JD Webster and RW Dunstan. Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology. *Veterinary Pathology*, 51(1):211–223, 2014.

- [226] Benzheng Wei, Zhongyi Han, Xueying He, and Yilong Yin. Deep learning model based breast cancer histopathological image classification. In *Cloud Computing and Big Data Analysis (ICCCBDA), 2017 IEEE 2nd International Conference on*, pages 348–353. IEEE, 2017.
- [227] Dominik Weishaupt, Victor D Köchli, and Borut Marincek. *How does MRI work?: an introduction to the physics and function of magnetic resonance imaging*. Springer Science & Business Media, 2008.
- [228] Nick Weiss, Henning Kost, and André Homeyer. Towards interactive breast tumor classification using transfer learning. In *International Conference Image Analysis and Recognition*, pages 727–736. Springer, 2018.
- [229] Jianxin Wu and Jim M Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.
- [230] Juanying Xie, Ran Liu, Joseph Luttrell IV, and Chaoyang Zhang. Deep learning based analysis of histopathological images of breast cancer. *Frontiers in Genetics*, 10:80, 2019.
- [231] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.
- [232] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, Jon Garibaldi, Ian O Ellis, Andy Green, Linlin Shen, and Guoping Qiu. Attention by selection: A deep selective attention approach to breast cancer classification. *IEEE Transactions on Medical Imaging*, 2019.
- [233] Ekin Yagis, Alba G Seco De Herrera, and Luca Citi. Generalization performance of deep learning models in neurodegenerative disease classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1692–1698. IEEE, 2019.
- [234] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- [235] Cuiru Yu, Houjin Chen, Yanfeng Li, Yahui Peng, Jupeng Li, and Fan Yang. Breast cancer classification in pathological images based on hybrid features. *Multimedia Tools and Applications*, 78(15):21325–21345, 2019.
- [236] EA Zanaty. Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, 13(3):177–183, 2012.
- [237] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

- [238] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.
- [239] Yuanjie Zheng, Sajjad Baloch, Sarah Englander, Mitchell D Schnall, and Dinggang Shen. Segmentation and classification of breast tumor using dynamic contrast-enhanced mr images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 393–401. Springer, 2007.
- [240] Weiming Zhi, Henry Wing Fung Yueng, Zhenghao Chen, Seid Miad Zandavi, Zhicheng Lu, and Yuk Ying Chung. Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images. In *International Conference on Neural Information Processing*, pages 669–676. Springer, 2017.
- [241] Shichong Zhou, Jun Shi, Jie Zhu, Yin Cai, and Ruiling Wang. Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. *Biomedical Signal Processing and Control*, 8(6):688–696, 2013.
- [242] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2079–2088, 2019.
- [243] Ziqi Zhu, Xinge You, CL Philip Chen, Dacheng Tao, Weihua Ou, Xiubao Jiang, and Jixin Zou. An adaptive hybrid pattern for noise-robust texture analysis. *Pattern Recognition*, 48(8):2592–2608, 2015.
- [244] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.