UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DEVELOPING NOVEL COMPUTER AIDED DIAGNOSIS SCHEMES FOR IMPROVED CLASSIFICATION OF MAMMOGRAPHY DETECTED MASSES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree Of

DOCTOR OF PHILOSOPHY

By

MEREDITH ANN JONES

Norman, Oklahoma
2023

DEVELOPING NOVEL COMPUTER AIDED DIAGNOSIS SCHEMES FOR IMPROVED CLASSIFICATION OF MAMMOGRAPHY DETECTED MASSES

A DISSERTATION APPROVED FOR THE
STEPHENSON SCHOOL OF BIOMEDICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Yuchen Qiu, Chair

Dr. Javier Jo

Dr. Qinggong Tang

Dr. John Antonio

Acknowledgements

The journey to complete my PhD has been a challenging and exciting rollercoaster. None of this would have been possible without the support of my incredible family, friends, and mentors.

I would like to express my deepest thanks to my mentors, Dr. Bin Zheng and Dr. Yuchen Qiu for showing me tremendous support and guidance over the past four years. I am deeply grateful to Dr. Zheng whose unwavering kindness and guidance have been a beacon of light throughout this academic journey. His mentorship went beyond mere academic advice; it extended to helping me find my way in the labyrinth of research, nurturing my growth as a critical thinker, and instilling in me the values of perseverance and resilience. His insightful feedback and constructive criticism were instrumental in refining my research and shaping my intellectual capabilities. I owe a deep sense of gratitude to Dr. Qiu for his willingness to mentor me throughout the final months of my journey. His time and support have allowed me to accomplish my goals of completing this dissertation.

I am extremely thankful for my committee members, Dr. Javier Jo, Dr. Qinggong Tang, and Dr. John Antonio for all their support and guidance in completing this dissertation. I want to thank my peers, Bill MaCcuaig, Rowzat Faiz, and Warid Islam, for their unwavering words of encouragement, support, and friendship. I also would like to thank Dean Randa Shehab and Dr. Michael Detamore for their incredible guidance throughout my PhD journey. They have been instrumental in helping me jump over any hurdle thrown my way, and for that I am forever grateful.

Lastly, it is difficult to put into words the gratitude I have for my family and friends. To my mom and dad, Brooke and Bobby, my sister, Allison, my brothers, Jack and Matt, my boyfriend, Corey, and my aunts, uncles, nieces, nephew, and cousins I thank you for all the love and support you have shown me over the past four years. It really does take a village, and I couldn't have done it without you all.

To my beloved mom, whose strength and resilience in the face of stage IV breast cancer inspired this research, this work is dedicated to you. Your courage fuels my commitment to making a difference.

Table of Contents

List of Tables

List of Figures

Abstract

Mammography imaging is a population-based breast cancer screening tool that has greatly aided in the decrease in breast cancer mortality over time. Although mammography is the most frequently employed breast imaging modality, its performance is often unsatisfactory with low sensitivity and high false positive rates. This is due to the fact that reading and interpreting mammography images remains difficult due to the heterogeneity of breast tumors and dense overlapping fibroglandular tissue. To help overcome these clinical challenges, researchers have made great efforts to develop computer-aided detection and/or diagnosis (CAD) schemes to provide radiologists with decision-making support tools. In this dissertation, I investigate several novel methods for improving the performance of a CAD system in distinguishing between malignant and benign masses.

The first study, we test the hypothesis that handcrafted radiomics features and deep learning features contain complementary information, therefore the fusion of these two types of features will increase the feature representation of each mass and improve the performance of CAD system in distinguishing malignant and benign masses. Regions of interest (ROI) surrounding suspicious masses are extracted and two types of features are computed. The first set consists of 40 radiomic features and the second set includes deep learning (DL) features computed from a pretrained VGG16 network. DL features are extracted from two pseudo color image sets, producing a total of three feature vectors after feature extraction, namely: handcrafted, DL-stacked, DL-pseudo. Linear support vector machines (SVM) are trained using each feature set alone and in combinations. Results show that the fusion CAD system significantly outperforms the systems using either feature type alone (AUC=0.756±0.042 $p<0.05$). This study demonstrates that both handcrafted and DL futures contain useful complementary information and that fusion of these two types of features increases the CAD classification performance.

In the second study, we expand upon our first study and develop a novel CAD framework that fuses information extracted from ipsilateral views of bilateral mammograms using both DL and radiomics feature extraction methods. Each case in this study is represented by four images which includes the craniocaudal (CC) and mediolateral oblique (MLO) view of left and right breast. First, we extract matching ROIs from each of the four views using an ipsilateral matching and bilateral registration scheme to ensure masses are appropriately matched. Next, the handcrafted radiomics features and VGG16 model-generated features are extracted from each ROI resulting in eight feature vectors. Then, after reducing feature dimensionality and quantifying the bilateral asymmetry, we test four fusion methods. Results show that multi-view CAD systems significantly outperform single-view systems (AUC = 0.876±0.031 vs AUC = 0.817±0.026 for CC view and 0.792±0.026 for MLO view, $p<0.001$). The study demonstrates that the shift from single-view CAD to four-view CAD and the inclusion of both deep transfer learning and radiomics features increases the feature representation of the mass thus improves CAD performance in distinguishing between malignant and benign breast lesions.

In the third study, we build upon the first and second studies and investigate the effects of pseudo color image generation in classifying suspicious mammography detected breast lesions as malignant or benign using deep transfer learning in a multi-view CAD scheme. Seven pseudo color image sets are created through a combination of the original grayscale image, a histogram equalized image, a bilaterally filtered image, and a segmented mass image. Using the multi-view CAD framework developed in the previous study, we observe that the two pseudo-color sets created using a segmented mass in one of the three image channels performed significantly better than all other pseudo-color sets (AUC=0.882, p<0.05 for all comparisons and AUC=0.889, p<0.05 for all comparisons). The results of this study support our hypothesis that pseudo color images generated with a segmented mass optimize the mammogram image feature representation by providing increased complementary information to the CADx scheme which results in an increase in the performance in classifying suspicious mammography detected breast lesions as malignant or benign.

In summary, each of the studies presented in this dissertation aim to increase the accuracy of a CAD system in classifying suspicious mammography detected masses. Each of these studies takes a novel approach to increase the feature representation of the mass that needs to be classified. The results of each study demonstrate the potential utility of these CAD schemes as an aid to radiologists in the clinical workflow.

Chapter 1. Introduction

1.1. Background

The latest cancer statistics data for the USA estimates that in 2022, 31% of cancer cases detected in women are breast cancer with 43,250 cases resulting in death. This accounts for 15% of total cancer-related deaths [1]. Thus, breast cancer remains the most diagnosed cancer among women with the second highest mortality rate. From 1989 to 2017, the mortality rate of breast cancer dropped 40% which translates to 375,900 breast cancer deaths averted [2]. Even though the mortality rate continues to decline, the rate of decline has slowed from 1.9% per year from 1998-2011 to 1.3% per year from 2011-2017 [2]. Over the past three decades, population-based breast cancer screening has played an important role in helping detect breast cancer in the early stage and reduce the mortality rate. However, the efficacy of population-based breast cancer screening is a controversial topic due to the low cancer prevalence (≤0.3%) in annual breast cancer screening resulting in a low cancer detection yield and high false-positive rate [3]. This high false positive rate is indicative of a high rate of unnecessary biopsies which is not only an economic burden but also leads to unnecessary patient anxieties which often result in women being less likely to continue with routine breast cancer screening [4].

Conversations pertaining to the benefits and harms of screening mammography as well as its efficacy in decreasing breast cancer mortality as screening exams do not reduce the incidence of advanced/aggressive cancers are now common [5]. For example, detection of ductal carcinoma in situ (DCIS) or early invasive cancers that will never progress or be of risk to the patient are occurring at a disproportionately higher rate than aggressive cancers. This is referred to as overdiagnosis and often results in unnecessary treatment that may cause more harm than the cancer itself [6]. Thus, improving the efficacy of breast cancer detection and/or diagnosis remains an extremely pressing global health issue [7].

While advances in medical imaging technology and progress towards better understanding the complex biological and chemical nature of breast cancer have greatly contributed to the large decline in breast cancer mortality, breast cancer is a complex and dynamic process, making cancer management a difficult journey with many hurdles along the way. The cancer detection and management pipeline has many steps, including detecting suspicious tumors, diagnosing said tumors as malignant or benign, staging the subtype and histological grade of a cancer, developing an optimal treatment plan, identifying tumor margins for surgical resections, evaluating and predicting response to chemo or radiation therapies, or predicting risk of future occurrence or reoccurrence. In this clinical pipeline, medical imaging plays a crucial role in the decision-making process for each of these tasks. Traditionally, radiologists will rely on qualitative or semi-quantitative information visually extracted from medical images to detect suspicious tumors, predict the likelihood of malignancy, and evaluate cancer prognosis. The clinically relevant image information may include enhancement patterns, presence or absence of necrosis or blood, density and size of suspicious tumors, tumor boundary margin spiculation, or location of the suspicious tumor. However, interpreting and integrating information visually detected from medical images to make a final diagnostic decision is not an easy task.

1.2. Computer-aided diagnosis (CAD) in mammography imaging

Mammography imaging is the most commonly used and widely available breast imaging modality. It plays a crucial role in the early diagnosis of breast cancer which is critical for keeping the chance of morality low. Obtaining an accurate diagnosis depends on the radiologists ability to accurately interpret the images. Mammography imaging uses low energy x-rays to create a 2D projection of the breast. Since it is an x-ray-based technique, the resulting images tend to be noisy and low contrast which make image interpretation difficult. In addition, the denser a breast is the more difficult it is for a radiologist to assess as abnormal regions may be obscured by overlapping dense tissue (**Figure 1-1**). This leads to a pressing clinical problem as dense breasts are associated with a greater risk of future breast cancer, yet it is much more difficult for a radiologist to interpret mammography images taken of a dense breast [8]. While diagnosing breast cancer from mammography exams is the most crucial step in the cancer management pipeline, the image interpretation task is a subjective process that can vary between radiologists. It is also extremely time intensive, which can lead to inaccurate results from fatigue associated with reading hundreds of images, and costly as two radiologists must review every exam and reach a consensus. If a consensus is not possible, a third radiologist must also review the exam to reach a decision[9].

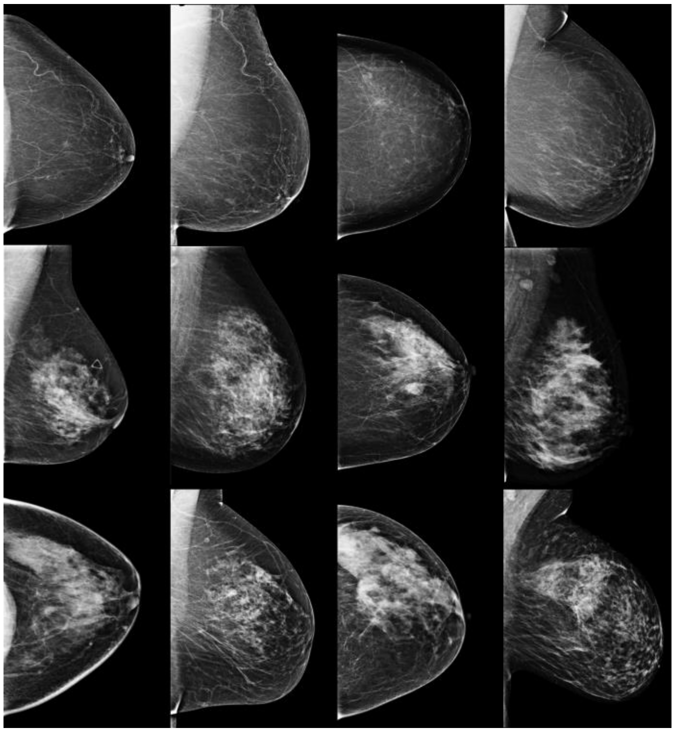

Figure 1-1: *Examples of different mammogram images. The top row represents normal density mammogram images. The bottom two rows depict the heterogeneity of fibroglandular tissues and dense tissues. Brighter regions correspond to areas of high density. These factors make it difficult to see if there is a lesion being obscured by the dense patches*

To address these clinical challenges, computer aided detection/diagnosis (CAD) schemes have been proposed to assist radiologists in more accurately and efficiently reading and interpreting medical images [10, 11]. CAD systems use quantitative image features to analyze a set of images and provide the radiologist with a second opinion or decision support to optimize the busy clinical workflow. The idea to use computers to automatically analyze mammography images dates back to the 1960s[12, 13]. While these studies demonstrated the feasibility of CAD systems, performance was overall poor due to computational limitations. In the 1980s and 1990s, CAD system development shifted from focusing on fully automated systems to systems that assisted radiologists as a second reader, resulting in more attention. Observer studies highlighted the undeniable potential of these CAD systems in breast cancer[14, 15]. In 1998, the FDA approved the first CAD system for mammography [16]. This system was widely adopted, one study reported that in 2016 CAD was used in about 92% of screening mammograms read in the United States [16, 17]. Despite this wide scale clinical adoption, the utility of commercialized CAD schemes for breast cancer screening is often questioned [18-20].

Since then, advances in artificial intelligence have demonstrated incredible capabilities in image analysis tasks such as detection, segmentation, and classification. This is part of the reason for the immense research interest in developing CAD systems for mammography over the past two decades. Despite this widespread attention, only eight CAD systems have received FDA approval for the detection and classification of breast cancer since the first system in 1998 (**Table 1-1**) [9]. Many of these systems have been approved within the last decade. This can be explained by the recent boom in deep learning techniques which has also diffused into CAD based mammography research.

Tool	**Company**	**Country**	**Application**	**Date of Approval**
cmTriage[21]	CureMetrix	United States	Triage	08/03/2019
HealthMammo[22]	Zebra Medical Vision	Israel	Triage	16/07/2020
Saige-Q[23]	DeepHealth	United States	Triage	16/04/2021
MammoScreen [24]	Therapixel	France	Detection and classification	25/03/2020
Genius AI Detection [25]	Hologic	United States	Detection and classification	15/11/2020
ProFound AI Software [26, 27]	iCAD	United States	Detection and Classification	12/03/2021

Transapra 1.7.0 [28]	ScreenPoint Medical B.V.	Netherlands	Detection and Classification	02/06/2021
INSIGHT MMG [29]	Lunit	Korea	Detection and Classification	17/11/2021

Table 1-1: *FDA approved CAD systems for mammography. Table taken from[9]*

In the literature, CAD is often differentiated as computer-aided detection (CADe) or computer-aided diagnosis (CADx). The goal of CADe schemes is to reduce observational oversight by drawing the attention of radiologists to suspicious regions in an image. On the other hand, the goal of computer-aided diagnosis (CADx) schemes is to characterize a suspicious area and assign it to a specific class. The work described in this dissertation focuses solely on CADx schemes for mammography. For this dissertation, CAD and CADx are used interchangeably. CAD schemes can be divided into two classes, traditional CAD or machine learning (ML) based CAD, and deep learning (DL) based CAD (**Figure 1-2**).

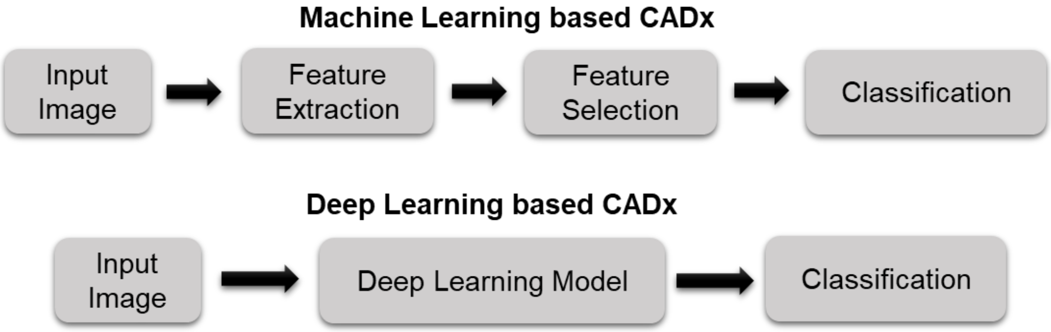

Figure 1-2: *Schematic diagram representing the foundational steps of ML and DL based CAD schemes.*

1.2.1. Machine learning based CAD

While DL-based CAD schemes have become increasingly more common than ML based CAD schemes, they still have their advantages over deep learning systems. In general, ML-based CAD systems are often considered more explainable than DL-based CAD systems as these models tend to provide more transparency into the decision-making process by following an explicit set of rules which can be explained. While DL-models have a “black-box” nature which makes it difficult to understand and interpret the rationale behind the decisions. Additionally, ML systems do not require nearly the amount of training data that DL systems do. Data limitations are unfortunately a common problem in the medical imaging domain. Recent attention to deep learning has pushed for the curation of large publicly available datasets which will continue to sway developers towards DL based systems. However, ML-based systems will be immensely useful in scenarios where this dataset is not available. ML-based CAD systems are also less computationally demanding, making them much more accessible as they can be deployed in resource-constrained environments.

ML-based CAD systems traditionally contain four main steps. First, a region of interest (ROI) is defined. Second, a set of quantitative features is extracted to characterize the ROI. Third, the feature set is reduced to an optimal feature set. And fourth, a classifier is trained and tested to predict the likelihood of the ROI being in either class. A brief explanation of each of these steps is found in the following sections.

1.2.1.1 Definition of a region of interest

In order to classify a suspicious tumor, the region in which to extract features must be defined. In mammography-based CAD schemes, this is most commonly the tumor or a

ROI surrounding the tumor. Accurate segmentation of the ROI is a very hot topic in the medical imaging domain [30, 31]. There are various ways to segment a desired region; these can be broken down into three broad groups: manual segmentation methods, semi-automated segmentation methods, or fully automated segmentation methods.

Manual segmentation involves a human annotating the boundary of each tumor or desired region from the background regions. This method allows for high precision as it is conducted by experts with full control, allowing for adjustments due to artifacts or irregularities. However, this is an extremely time intensive task leading to fatigue which may decrease the accuracy of segmentation. Manual segmentation is also subjective, meaning multiple radiologists may annotate the same image differently leading to unwanted variability in the image inputs. Semi-automated segmentation methods use a combination of human input and computer-based algorithms to get the segmentation results. This could be in the form of a human using prior knowledge of the tumor location to set an initial seed, or a human manually correcting the segmentation result. In the literature, many CADx studies rely on the semi-automated method of extracting a ROI of a fixed size surrounding the center of each tumor that has been marked by a human. Automated segmentation schemes are able to conduct image segmentation without any human intervention. Fully automated methods are advantageous in that they are much more time efficient and consistent than the other methods, but the performance of these methods is dependent on the segmentation task meaning they may perform poorly in ill-defined or difficult segmentation tasks.

The segmentation task is non-trivial as there are many factors that make it difficult. This includes dense breast tissue that may obscure the mass boundary, and the pectoral muscle or artifacts and distortions which may fool the segmentation algorithms. Over the years, various techniques have been proposed for the segmentation of ROIs from mammograms. These techniques range from classical segmentation methods that rely on pixel intensities, to machine learning and deep learning segmentation methods which require model training and testing[30, 32]. There is no consensus on the best segmentation method to use to extract a ROI for mammography based CADx schemes. In addition, there is also no consensus on the best location to extract an ROI from. For example, some studies may opt to extract a bounding box that surrounds the tumor[33], other studies may choose to use the tumor boundary[34], while other studies may use the whole breast image[35, 36]. The decision is based on the goal of the study. For example, studies focused on predicting if a tumor is malignant or benign will tend to focus more on the tumor and surrounding tissue, while studies focusing on predicting the risk of breast cancer in breast images without tumors may focus on the area behind the nipple as it most accurately describes the breast parenchyma patterns which is an established biomarker of breast cancer risk[37].

Overall, tumor segmentation remains one of the most difficult challenges that traditional ML based CADx schemes encounter and a major hurdle to clinical implementation. The shift from manual to semi-automated to fully automated lesion segmentation has decreased the inherent bias associated with human intervention, but elimination of the segmentation step in its entirety through CNNs will allow for more generalizable CADx systems.

1.2.1.2 Feature extraction

After a ROI is defined, a set of handcrafted radiomic features will be extracted. While using radiological features from medical images to infer phenotypic information has been done for many years, recent rapid advances in bioinformatics coupled with the advent of high performing computers has led to the field of radiomics. Radiomics involves the transformation of images into mineable data through the computation of quantitative image-based features. These features can then be leveraged in clinical decision support systems to predict clinical outcomes and tailor treatment planning to individual patients, further shifting us towards the new paradigm of personalized medicine [38, 39].

Feature extraction is the most crucial step in developing ML-based CADx schemes as the feature set will be the input to the classification model. If the feature set does not accurately capture the characteristics of each class, then the model will not be able to learn sufficiently and thus have a poor performance.

One of the main advantages of handcrafting a feature set is that it benefits from domain knowledge meaning image characteristics that are known to be relevant to the task can be quantified and used as features. For example, malignant tumors as seen on mammograms are typically irregular in shape with spiculated margins and architectural distortions while benign tumors are typically rounded with well-defined margins **(Figure 1-3)** [32, 40, 41]. Quantification of these features can help train robust ML classifiers to better differentiate between benign and malignant masses. Features that describe the shape of the tumor may include eccentricity, diameter, convex area, orientation, and more[32]. Features can also be extracted to quantify the spiculations of the tumors which will be particularly helpful for detecting malignant breast tumors [42].

Other examples of common features are first order statistical features which describe the distribution of intensities within an image, this includes mean, standard deviation, variance, entropy, uniformity, and others. Entropy quantifies the image histogram randomness which can quantify heterogeneity of the image patterns [43]. Texture features belong to the biggest group of radiomics features, which are extremely useful for image recognition and image classification tasks [44, 45]. Gray-level cooccurrence matrix (GLCM) based features and gray-level run length matrix (GLRLM) based features are two example of common texture features that characterizes the heterogeneity of intensities within a neighborhood of pixels. Quantification of the heterogeneity of tumors is one of the advantages of radiomics-generated imaging markers as heterogeneity is often very difficult for radiologists to visually capture and quantify in clinical practice.

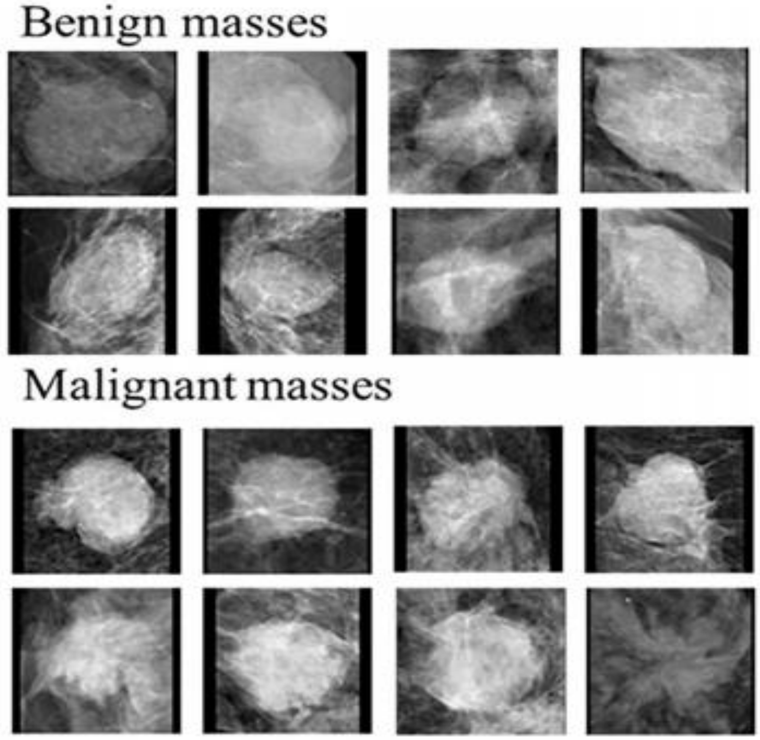

Figure 1-3: *Examples of malignant and benign masses seen on mammograms. Modified from [41].*

While there has been a wide variety of radiomics features extracted from many different locations for different cancer applications, there is no consensus on what features make up an optimal feature set. Deciding what features should be extracted remains dependent on the goal of the individual study.

1.2.1.3 Feature selection and reduction

The initial feature set extracted from the ROI is often large and contains many highly correlated and irrelevant features that may decrease model performance if included in the final feature set. Additionally, ML-based CADx models are subject to the curse of dimensionality which asserts that after a certain point, the amount of data samples needed to train a machine learning classifier increases exponentially with a large number of input features[46]. Therefore, creation of an optimal feature subset from the initial feature pool is a critical step in building ML models.

The goal of feature selection and reduction is to identify a subset of the features that will yield the best model performance. Feature selection aims to identify a subset of the original features that are most relevant to the task. The final set of features selected also exists in the original feature set, meaning the features are not changed. Feature selection methods can be broken down into three categories: wrappers, filters, and embedded methods[47]. Wrapper methods use the classifier to drive the feature selection process by assessing multiple subsets of features effect on model performance. Some examples of these methods are exhaustive, greedy, or stochastic search algorithms, and sequential forward or backward selection algorithms. These methods tend to be very computationally expensive as these algorithms must search variations of all possible feature combinations. It often may be more practical to use these methods in combination with a filter method. Filter methods rank each feature based on a specific criterion that quantifies the importance of the feature in the prediction task then select the best features based on a selection criterion. Examples of filter methods commonly used are variance thresholding, correlation-based feature selection, and relief-based algorithms. Embedded methods are similar to wrapper methods in that feature importance is deduced during model training. These methods include L1-regularization and tree-based models like decision trees or XGBoost in which an optimal feature set can be deduced after the model has been trained[48].

Feature reduction methods are also commonly used in ML-based CADx schemes. These techniques differ from feature selection methods as they transform the initial high-dimensional feature set into a lower-dimensional representation. The main examples of this method are principal component analysis (PCA) which aims to capture the variance in the initial feature space while reducing the dimensionality, and linear discriminant analysis (LDA) which aims to reduce the initial feature dimensionality while maximizing the between-class separation and minimizing the within-class separation[49].

1.2.1.4 Classification

Machine learning classifiers are able to learn patterns within the input feature sets and classify an image as malignant or benign. ML techniques can either be unsupervised, supervised, or semi-supervised. In unsupervised learning, the algorithm explores patterns and structures within the data without labeled examples to guide it. These methods are helpful in identifying hidden structures and patterns in the data when labeled data is not available. Supervised learning leverages labeled data to train models, offering high accuracy and interpretability. However, it relies on a substantial amount of labeled data and may struggle with unknown classes. Semi-supervised learning combines elements of both, where a portion of the data is labeled while the rest remains unlabeled. It allows for efficient utilization of available labeled data and can generalize well but may still require a reasonable amount of labeled data for the best results. Each method has its place depending on the data and the problem at hand, making them versatile tools in the field of machine learning.

Supervised learning algorithms are most commonly used in mammography-based CAD systems. The most common ML classifiers used in mammography-based CAD systems are support vector machines (SVMs), artificial neural networks (ANNs), and K-nearest neighbors (KNNs) [50, 51]. Each ML method can be thought of as a mathematical model that takes a set of features and the labels that represent each image, and outputs a prediction. In the training phase, the predicted value is then compared to the true label via a loss function. The loss function measures how well the model is performing by quantifying the dissimilarity between the actual value and the predicted value, then adjusts the model parameters in a way that minimizes the loss function. This adjustment is done according to an optimization algorithm. Once the model is trained, it will then be tested using an independent image set. For the model to be considered a good classifier, it must perform well on not only the training set but also the testing set. This would indicate that the model can generalize well on unseen data and has truly learned.

1.2.2. Deep learning based CADx

Recent enthusiasm for deep learning (DL) based AI technology has led to new approaches for developing CAD schemes which are being rapidly explored and reported in the literature [52]. DL based CAD schemes use convolutional neural networks (CNNs) to automatically learn hierarchical representations of the images directly from the image, eliminating the need for semi-automated or fully automated tumor segmentation and handcrafted feature selection. CNNs use convolutional layers that apply filters (kernels) to local regions of the input image, allowing them to capture local patterns and features. This architecture is well-suited for grid-like data, like images, where spatial relationships are important. The selected filters and convolutional layers are what make the CNN a powerful tool as it enables it to detect, learn, and recognize different image features or patterns.

Briefly, the convolutional layers extract patterns from an input image by convolving the input image with a filter or kernel of specific weights. Patterns are organized into a feature map which will go through an activation function and be passed to the next layer. Without activation functions these networks would only be capable of linear feature mapping which would make it nearly impossible to learn features of complex non-linear distributions [53, 54]. Following the convolutional layers are subsampling layers, often max pooling layers, which will down sample the feature map by calculating the maximum value in a region. This highlights the most present feature in the map while decreasing the number of parameters that the model needs to learn and increasing the robustness [55]. One or more convolution layers followed by a pooling layer is often referred to as a block. Stacking multiple blocks is what makes this a deep network. The repetitive layer structure of the deep CNN is what allows for the extraction of increasingly meaningful information while preserving spatial information. The final feature map will be passed to one or more fully connected layers. The fully connected layers have full connections to all neurons in the previous layer allowing it to identify relationships between all features in the feature map and output a class prediction.

The main limitation of DL-based CADx schemes is the need for a large and diverse dataset to properly train the network. This is not often available in the medical imaging domain. Researchers have trained shallow CNNs for breast mass classification which do not require as much training data as a deep CNN model, but the robustness of these schemes is questionable as they are trained on smaller dataset [56-58]. The deeper a model is, the more complex representations can be learned, so the question of how deep a CNN must be to sufficiently capture features for a large classification task remains [59]. However, training a deep CNN from scratch is not possible without a large diverse dataset which are often not readily available in the medical imaging field.

By recognizing the limitation of shallow CNN models, transfer learning has emerged as a solution to lack of big data in medical imaging. In transfer learning, a CNN is trained in one domain and applied in a new target domain [60]. This involves taking advantage of existing CNNs that have been pretrained on a large data set like ImageNet and repurposing them for a new task as this allows the networks to obtain a good sense of computer vision [61, 62]. Thus, several well established state of the art CNNs such as AlexNet, GoogLe-Net, ResNet, VGG16, and others have been pre-trained on the ImageNet dataset and successfully used in a wide variety of computer vision tasks including detection, segmentation, and classification of medical images [61, 63, 64]. There are two approaches to transfer learning **(Figure 1-3).** Fine tuning involves freezing some layers of a pre-trained model while training other layers[65]. Feature extraction via transfer learning involves using a pre-trained network exactly as is to extract feature maps that will be used to train a separate ML model or classifier. The former is beneficial in that it will train the network to have some target specific features, but the latter is advantageous in that it is computationally inexpensive as it does not require any deep CNN training [66].

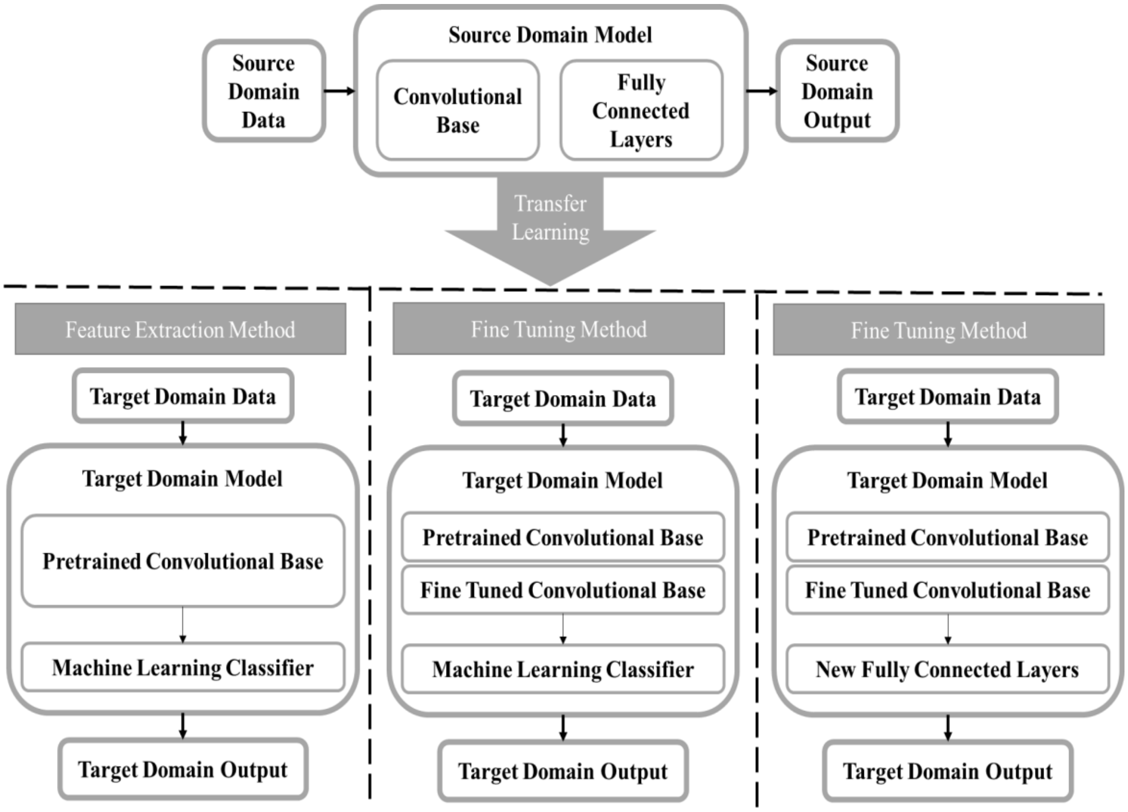

Figure 1-4: *A block diagram displaying the transfer learning process. A model is trained in the source domain using a large diverse dataset. The information learned by the model is transferred to the target domain and used on a new task. The two main methods for transfer learning are feature extraction and fine tuning. For the feature extraction method, a feature map is extracted from the convolutional base taken from the source model and used to train a separate machine learning classifier. There are two ways to use transfer learning by fine tuning. The first is freezing the initial layers in the convolutional base from the source model and fine tuning the final layers using the target domain dataset then training a separate classifier. The second method does the same, except instead of training a new machine learning classifier, new fully connected layers will be added and trained using the target domain data [67].*

1.3. Challenges of current CAD development

Despite the extensive research efforts dedicated to the development and testing of new AI-based models in the laboratory environment, very few of these studies or models have been translated into clinical practice. This can be attributed to several obstacles or challenges.

First, currently, most of the studies reported in the literature trained AI-based models using small datasets (i.e., <500 images). Training a model using a small dataset often results in poor generalizability and poor performance due to unavoidable bias and model overfitting.

Second, medical images acquired using different machines made by different companies and different image acquisition or scanning protocols in different medical centers or hospitals may have different image characteristics (i.e., image contrast or contrast-to-noise ratio). CAD schemes are often quite sensitive to the small variations of image characteristics due to the risk of overtraining. Thus, models developed in this manner are not easily translatable to independent test images acquired by different imaging machines at different clinical sites. Developing and implementing image pre-processing algorithms to effectively standardize or normalize images acquired from different machines or clinic sites [68, 69] have also attracted research interest and effort.

Third, as mentioned previously, another common limitation of traditional ML or radiomics based models is that they often require a lesion segmentation step prior to feature extraction. Whether lesion segmentation is done semi-automatically based on an initial seed or automatically without human intervention, accurate and robust segmentation of breast lesions from the highly heterogeneous background tissue remains difficult [70]. The lesion segmentation error introduces uncertainty or bias to the model due to the variation of the computed image features and hinders the translation of the models to the clinic. Recent attention to DL technology provides a way to overcome this limitation as the deep CNNs will extract features directly from the images themselves, bypassing the need for a lesion segmentation step. However, the lack of big and diverse datasets is a major challenge in developing robust DL models. Although transfer learning has emerged as a mainstream in the medical imaging field, its advantages and limitations are still under investigation. For example, while there is a huge focus on using pre-trained CNNs as feature extractors as it is computationally inexpensive and generalizable since these models avoid having to train or re-train the CNN at different centers with different imaging parameters, fine tuning the models has showed better results [60]. Additionally, no CNN-based transfer learning models have made it to clinical use since the models are still not robust as investigated in a recent comprehensive evaluation study [71]. Therefore, more development and validation studies are needed to address and overcome this challenge.

Fourth, currently most DL-based models use a “black-box” type approach and lack explainability. As a result, it reduces the confidence or willingness of clinicians to consider or accept AI-generated prediction results [72]. Understanding how the model can make reliable prediction is non-trivial to most individuals because it is very difficult to explain the clinical or physical meanings of the features automatically extracted by a CNN-based deep transfer learning model. Thus, developing explainable models in medical image analysis has emerged as a hot research topic [73]. Among these efforts, visualization tools with interactive capability or functions have been developed that aim to show the user what regions in an image or image patterns (i.e., “heat maps”) contribute the most to the decision made by a models [74, 75]. In general, new explainable AI models must be able to provide sound interpretation of how the features extracted result in the output produced. Ideally this should be done in ways that directly tie to the medical condition in question. Since this is an emerging research field and important research direction, more research efforts should dedicate to extensive development of new technologies to make CAD schemes and prediction models more transparent, interpretable, and explainable before the AI-based models or decision-making supporting tools can be fully accepted by the clinicians and then integrated into the clinical workflow.

Chapter 2. Research Objectives and Hypothesis

It is important to investigate new methods to help decrease the false positive recall and benign biopsy rates of mammography so that women continue participating in routine breast cancer screening. As described in chapter 1, to help overcome these clinical challenges, researchers have made great efforts to develop CAD schemes for mammography to provide radiologists with decision-making support tools. Despite vast research efforts, the added clinical value is limited. Thus, more novel research efforts are needed to explore new approaches [76].

The overall objective of this dissertation is to investigate three unique research ideas for improving the performance of CAD schemes in classifying suspicious mammography masses based on current gaps in the literature. This includes (1) investigating the advantages of the fusion of traditional radiomics features typically used in ML-based CAD frameworks with deep learning-based features, (2) developing a novel multi-view CAD framework that uses true case-based inputs, and (3) investigating the role of pseudo color image generation in increasing mammography image feature representation prior to classification. The motivations, hypothesis, and proposed approach for each of these three studies are briefly discussed in the following section.

2.1. Feature level fusion of radiomics features and DL features to improve lesion classification

2.1.1. Background and motivations

Extensive research has demonstrated the potential of both ML-based CADx schemes and DL-based CADx schemes to improve the accuracy of classifying suspicious mammography detected masses as malignant or benign. [77, 78]. In the previous chapter, we discussed the strengths and limitations of both traditional ML and DL techniques and noted that where one technique fails, the other may succeed. For example, handcrafted features can closely mimic image features or markers used by radiologists in lesion diagnosis, while automated features can extract new clinically relevant features that may be invisible to the human eye. This has led to recent attention in developing CAD schemes that take advantage of both traditional and DL-based methods [41, 79, 80]. However, this is a new approach and further work must be done to identify optimal methods for fusing information from these techniques. The objective of this study is to develop a new CADx framework to effectively fuse a handcrafted radiomics feature set created using domain knowledge, and a DL feature set created using transfer learning techniques, to improve mass feature representation and improve the classification performance.

2.1.2. Hypothesis and proposed approach

In this study, we investigate the hypothesis that traditional handcrafted radiomic features and deep learning model generated features contain complementary discriminatory information and the fusion of these two types of features can increase the performance of a CADx system in classifying malignant and benign breast lesions. From a ROI surrounding suspicious mammography-detected masses, a handcrafted radiomics and a DL generated feature set is extracted. Due to the limited size of the dataset used, the DL features are extracted via transfer learning. The VGG16 network pretrained on the ImageNet database is used in this study to extract an extremely large feature pool. The images input to the DL network must be transformed from single-channel greyscale images into three-channel images to match the ImageNet dataset that the network is pretrained on. We created two pseudo-color image sets. A stacked image set that contains the original ROI in three channels, and a pseudo image set which contains the original ROI, a bilaterally filtered version, and a histogram equalized version stacked in three channels. DL features are extracted from both the stacked and pseudo image sets, resulting in three extracted feature vectors. A novel feature reduction pipeline is used to reduce the dimensionality of the three feature sets. This allows us to also investigate the effects of including preprocessed variants in the DL feature extraction pipeline. The optimal radiomics and DL-generated feature sets are then concatenated together, creating two fusion feature vectors. The five final feature vectors are then used to train a SVM to classify suspicious lesion as malignant or benign.

2.2. A multi-view CADx framework for breast lesion classification

2.2.1. Background and motivations

Mammography imaging exams traditionally take two different projection images of each breast. A craniocaudal (CC) view is taken from topdown, while the mediolateral oblique (MLO) view is taken at an angle from the left to right. Radiologists will use all four images (left and right CC and MLO) to decide if a lesion is present and if that lesion is malignant or benign and needs to be biopsied. However, most CAD schemes are single view-based schemes, which limits the performance and clinical utility. To increase the performance of these CAD systems, there has been recent attention on multi-view information fusion of the different views which provide the CAD system will a better understanding of the cancer as more information is available. These systems tend to outperform single view systems, but it is still a relatively new research area (**Table 2-1**).

There is no consensus on the optimal method to extract and fuse information from all four views to build a multi-view CAD system.

We can divide the current multi-view CAD systems into three categories based on the images used as an input. First, using only ipsilateral views which extracts and fuses information from the CC and MLO images of one breast. Using ipsilateral views is advantageous since image characteristics that can help classify a lesion may be obscured by dense overlapping tissue in one view but can be fully visible in the other view. Second, using only bilateral views which uses the same projection image taken of both the left and right breast. Using bilateral views is advantageous as it allows for quantification of bilateral asymmetry, a well-established image-based biomarker of a breast abnormality. When a radiologist interprets screening mammogrpahy exams, bilateral asymmetry is used a qualitative indicator of abnormalities as locations of high asymmetry often contain a suspicious mass[81]. And third, using both ipsilateral and bilateral views which provides the maximum amount of information to the CAD systems but must be handled carefully.

When using all four images as a simultaneous input for a CAD scheme, careful consideration must be made to ensure that each image corresponds to the same mass and location, therefore it is truly a case-based scheme. One major limitation of existing multi-view studies is that many of them are not truly case-based, meaning the left and right CC and MLO view images used as an input may not be appropriately matched (**Table 2-1**). For example, many studies that use ROIs extracted from ipsilateral views do not match the lesions ipsilaterally while building the image set. It is very possible that there will be more than one suspicious mass observed in a breast, therefore it is important to ensure that the ROI extracted from the CC and MLO images are representative of the same mass.

The same logic applies when using ROIs from bilateral views. Mammography exams compress the breast tissue, therefore without proper image registration, the bilateral ROIs do not actually correspond to the same region. The anatomical deformations present in the compressed breast image make simple rigid or affine transformation techniques an improper choice for the bilateral image registration task. This often causes researchers to stray away from the registration task and either quantify bilateral image characteristics from mismatched breast regions or use whole breast images which may hinder classification sensitivity[35, 82].

Year	Author	Image Set	Views	Ipsilateral Matching?	Bilateral Registration?	Fusion Method	Model	Metric
2015	Tan et al. [35]	1896 private FFDM	Ipsilateral + bilateral	N/A - whole breast images used		bilateral features are concatenated into a final CC and MLO feature vector	multi-stage ANN	AUC: 0.779 ± 0.025
2019	Li et al.[83]	182 private FFDM	bilateral	N/A	No	features are extracted from each view independently then concatenated into a single vector	Bayesian artificial neural network	AUC: 0.84 ± 0.03
2019	Khan et al.[84]	CBIS-DDSM	Ipsilateral + bilateral	No	No	features are extracted from each view independently then concatenated into a single vector	VGGNet pretrained on ImageNet	AUC: 0.84
2020	Hina et al[85].	CBIS-DDSM mini-MIAS	Ipsilateral + bilateral	No	No	prediction scores of each view are generated independently and fused using an attention based weighted algorithm	ResNet50 pretrained on ImageNet	AUC: 0.896
2017	Geras et al. [86]	886,437 private FFDM	Ipsilateral + bilateral	N/A - whole breast images used		features are extracted from each view independently then concatenated into a single vector	CNN	macro-AUC: 0.733

2023	Boudouh et al.[87]	CBIS-DDSM	Ipsilateral	No	N/A	Ipsilateral features are concatenated	InceptionResNetV2 and EfficentNetB7 pretrained on ImageNet used to extract features from the CC and MLO views, respectively. Final FC layers are trained	ACC: 95.86%

Table 2-1: *Examples of studies that include multiple mammography views as an input to a CAD scheme focused on classifying masses as malignant or benign.*

Previous studies demonstrate that multi-view CAD models tend to outperform single-view CAD models[88], the addition of information from the contralateral breast to quantify the bilateral asymmetry increases model performance[89-91], and the fusion of handcrafted radiomic features and deep learning features outperforms either method alone when classifying suspicious breast lesions[41, 79]. However, to the best of our knowledge, no existing work combines these three points into a singular framework. Our previous study (section 2.1) used only craniocaudal view images as an input to the CAD system. In this study, we build upon our previous work to create a true case based multi-view CADx system that takes advantage of the bilateral asymmetry using both radiomics and DL features.

2.2.2. Hypothesis and proposed approach

We hypothesize that a four-view image-based CAD scheme can yield a significantly higher mammography detected tumor classification performance than a one-view or two-view system when using both handcrafted radiomics features and deep learning features.

This work contains a multi-stage fusion problem as we must fuse information from multiple images and two different feature types. We investigate multiple fusion methods to determine the optimal method to do so. The novelty of this work lies in the method in which we extract matching ROIs, making this a true case-based system. Each case contains four images, a left and right CC and MLO image, however the mass can only be seen in two of the four images. We first conduct an ipsilateral matching scheme using the images that contain the mass with the goal of confirming that the mass seen in ipsilateral views is the same mass. After ipsilateral matching, we register the breast images bilaterally so that matching ROIs can be extracted from the contralateral breast. A set of handcrafted radiomics features and DL features generated using a VGG16 network pretrained on ImageNet are extracted from all four images. We then quantify the bilateral asymmetry and then investigate multiple fusion methods to determine the optimal way to fuse information from multiple views using multiple feature types.

2.3. Investigating the effects of pseudo color image generation in classifying malignant and benign breast lesions

2.3.1. Background and motivations

Transfer learning techniques continue to be a useful and effective tool for creating systems that do not have a large and diverse training dataset available. Leveraging pre-trained models on datasets like ImageNet requires some network manipulation and image transformations prior to training as there are crucial differences between the ImageNet dataset and mammography images [62]. The ImageNet dataset is comprised of natural color (RGB) images that often have a singular distinct focal point to direct the classification, while mammograms are single channel greyscale images that appear relatively homogenous in comparison to the ImageNet dataset (**Figure 2-2**). Since ImageNet contains 3-channel color images while mammography images are single channel greyscale images, mammography images must be transformed into three-channel images to be suitable for transfer learning with ImageNet. We call this step pseudo color image generation.

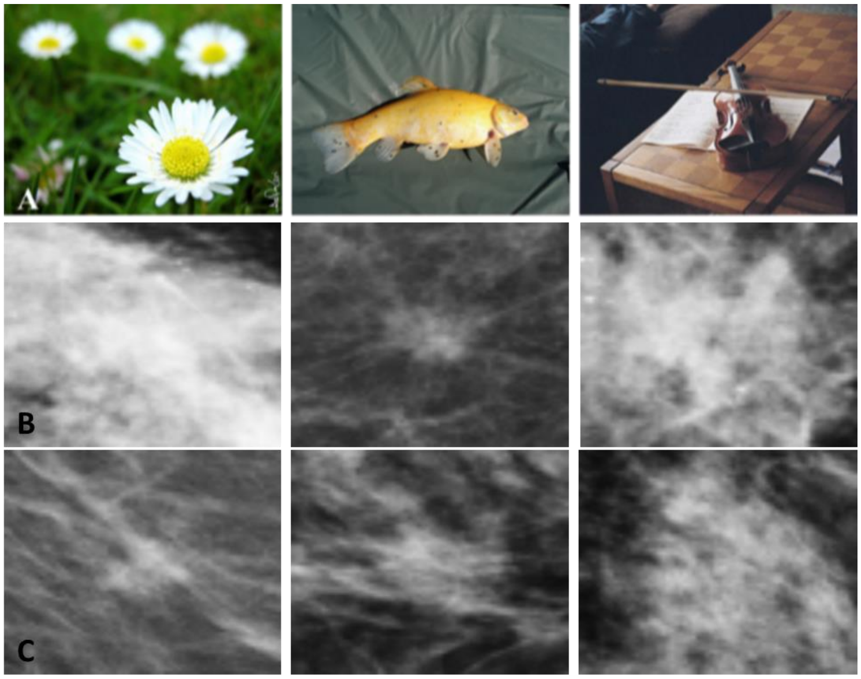

Figure 2-1: *Row A contains examples of the 3-channel color images in the ImageNet database. Row B and C contain single channel greyscale mammogram ROIs. Malignant cases are in row B while benign cases are in row C.*

In my first study mentioned in section 2.1, we extracted DL features from two pseudo color image sets. The first image set contained the original ROI stacked in three channels. The second image set contained the original ROI in the red channel and preprocessed variants in the green and blue channels, namely: a histogram equalized image and a bilaterally filtered image. The image set created using preprocessed variants performed better than the set created using only the original ROI, however this was only statistically significant in one of four comparisons. This has motivated us to continue the investigation into the effects of pseudo color image generation in deep transfer learning-based CAD schemes.

As I continued my literature search on the topic, I noticed that many studies omit details pertaining to the inputs fed to deep pretrained CNN networks for breast cancer. We believe it is assumed that the original image is stacked in 3-channels. Additionally, there are a very limited number of studies that investigate the effects of using pseudo color images in mammography mass classification and detection tasks (**Table 2-2**).

Year	**Author**	**Dataset**	**Task**	**Pseudo Color Image Generation**			**Model**	**Evaluation Metrics**
				R	**G**	**B**		
2017	Antropova et al.[79]	739 private FFDM	classification	original image	original image	original image	radiomics and DL features train separate SVMs the output scores are averaged	AUC: 0.86
2023	Razali et al.[92]	Inbreast	classification	parula colormapped			ResNet50 pretrained on ImageNet	AUC: 0.97
2017	Teare et al.[93]	DDSM and ZMDS	classification	CLAHE (window=2 clipping=8)	CLAHE (window=4 clipping=4)	CLAHE (window=8 clipping=2)	InceptionV3 pretrained on ImageNet as a feature extractor followed by a random forest	AUC: 0.92
2018	Li et al.[94]	352 private FFDM	detection	original image	gradient image	local ternary pattern	CNN	values not provided
2020	Min et al.[95]	Inbreast	detection and segmentation	original image	MMS image	MMS image	pretrained Mask R-CNN	TPR: 0.9 DSC: 0.88

Table 2-2: *Examples of mammography-based CAD studies that use a pseudo color image generation step.*

Notably, Li et al. created a pseudo color image set by stacking the original mammography ROI, a gradient image, and the local ternary pattern image for breast mass detection. There were no statistically significant differences in the ability of the system to detect suspicious masses when using the pseudo color image set compared to the single channel greyscale image. The authors speculate that this is because the variant images in the green and blue channels of the pseudo color image are created by convolving a kernel over the image, therefore the CNN may be able to learn these details on its own so including them in a pseudo color image is not actually increasing the feature representation [94]. Min et al. followed this work and created pseudo color images that added increased morphological information which CNN would not be able to learn on its own. The morphological information is in the form of images created from the original ROI using a multi-scale morphological sifter (MMS). Using pseudo color ROIs with increased morphological information outperformed the original ROIs in a mammography mass detection task[95].

It is well known that there are distinct morphological differences between malignant and benign tumors as malignant tumors tend to appear irregular in shape with spiculated margins while benign tumors appear round with defined boundaries (**Figure 2-2**). Many radiomics features are engineered using this domain knowledge to quantify these characteristics, but it is unclear how they are accounted for when using CNNs due to the black box nature. Additionally, the results of our first study (chapter 3) demonstrate that the domain knowledge used to build a handcrafted radiomics feature set is useful and should not be fully ignored. Since DL systems often lack any kind of domain knowledge, we believe adding morphological characteristics to a pseudo color image prior to using a pretrained CNN may increase the feature representation therefore yield better classification results.

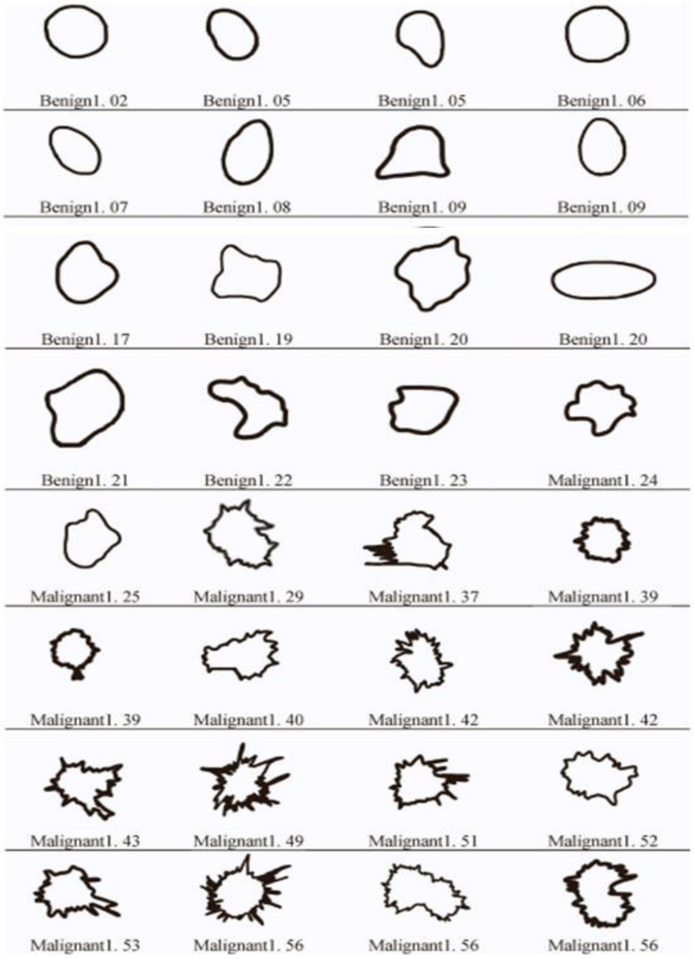

Figure 2-2: *Morphological examples of malignant and benign lesions. Modified from [30].*

2.3.2. Hypothesis and proposed approach

The purpose of this study is to continue the work from our previous two studies by fully investigating the effects of pseudo color image generation in classifying suspicious mammography detected breast lesions as malignant or benign using deep transfer learning in a multi-view CAD scheme. The performance of seven pseudo color image sets is compared. Pseudo color sets are created through combination of the original grayscale image, a histogram equalized image, a bilaterally filtered image, and a segmented mass image. We hypothesize that creating pseudo color images with additional morphological information will provide increased complementary information to a deep network pre-trained on the ImageNet database, and that this will yield better performance in classifying malignant and benign lesions than when using pseudo color images that do not contain morphological information.

To create pseudo color image sets with increased morphological information, a fully segmented mass image is used in one of the three channels. As mentioned in chapter one, the mass segmentation task is extremely non-trivial. Two different techniques are used to obtain this segmentation. The first is a manual segmentation where the mass boundary is drawn by hand. This is an extremely time-consuming, error prone, and subjective task that is rarely conducted in clinical practice[96]. To combat this limitation, we also generate a fully segmented mass using a Unet that uses the manual segmentation images as a ground truth. The idea is to demonstrate the feasibility of using an automated segmentation method to generate morphological information, so the manual segmentation task does not need to be added to the clinical workflow.

The CAD framework used in this study is taken from the study mentioned in section 2.2 as multi-view CAD always outperforms single view CAD. This is a true case-based system, as an ipsilateral matching and bilateral registration scheme is conducted in the same manner.

Chapter 3. Improving Mammography Lesion Classification by Optimal Fusion of Handcrafted and Deep Transfer Learning Features

3.1. Introduction

Breast cancer has the highest incident rate and second highest mortality rate among cancers in women [97]. Routine mammographic screening is considered a widely used cost-effective approach to detect breast cancer in its earliest stages, which can help significantly improve cancer treatment efficacy and reduce patients' mortality rate as demonstrated in many clinical studies [98, 99]. While mammography is the only accepted population-based breast cancer screening tool currently in clinical practice, mammograms are often difficult for radiologists to interpret due to the great heterogeneity of breast lesions and overlapped dense fibro-glandular tissues, which results in a high false positive recall rate. Among the suspicious breast lesions detected in mammograms and recommended for biopsy by radiologists, less than 30% of lesions are actually confirmed as malignant [100]. The high rate of benign biopsies is not only an economic burden, but also results in long-term psychosocial consequences to many women who participate in mammography screening [101]. Thus, improving the accuracy of classifying mammography-detected suspicious lesions to reduce the false-positive recall rate is a pressing clinical challenge.

One method to help improve breast lesion detection and classification, and the accuracy of radiologists is through the assistance of computer-aided detection and/or diagnosis (CAD) schemes. Typically, computer-aided detection schemes are developed and applied to detect and highlight locations of suspicious lesions depicting on mammograms, which may end up overlooked by radiologists, thus help increase lesion detection sensitivity (or reduce false negative rate) [102]. In addition, many other researchers have focused substantial efforts on the development and clinical translation of computer-aided diagnosis schemes that aim to classify the suspicious lesions as malignant or benign. In this article, we only develop and discuss computer-aided diagnosis (CAD) schemes. All CAD schemes include machine learning classifiers trained using a set of optimal image features extracted using one of two approaches. The first approach, often referred to as traditional CAD, involves extraction of a set of handcrafted radiomics image features to train a machine learning classifier. However, previous research indicates that extraction and selection of a set of optimal handcrafted features varies drastically between studies and is a time intensive, error-prone, and non-trivial task which often leads to increased false positive rates [43, 103, 104].

In order to overcome the challenges or limitations of the traditional CAD, many researchers investigated a second approach that uses a deep learning model to automatically learn and extract features directly from the image itself, which significantly decreases or eliminates user intervention. The deep learning models applied to medical images are primarily the deep convolution neural networks (CNN) due to their immense success in many tasks involving computer vision. CNNs differ from traditional artificial neural networks (ANNs) in that they use filters and convolutional operations to transfer all information from neurons in one layer to the neurons in the next hidden layers, which are called convolutional layers. The selected filters and convolutional layers are what make

CNNs a powerful tool as it enables it to detect, learn, and recognize different image features or patterns.

While deep CNNs have become an immensely powerful tool for many different image classification tasks, there are several limitations that hinder their applications to medical imaging tasks. Firstly, many of these deep learning algorithms are thought of as a black box. This is a key weakness and hurdle when trying to translate these technologies to the clinic. Visualization techniques have been proposed which give insight into the type of features extracted from each convolutional layer [105-107]; the goal of these techniques is to provide some level of explanation for their decision-making process. Second, training of these deep neural networks requires a very large dataset, which is often not available in medical imaging. Transfer learning has emerged as a solution to this problem. Transfer learning involves the transfer of knowledge from one task to another by using the model learned on one task for a separate task [62]. The theory is that if the original model is trained on a very large and diverse dataset, then the model will have a good sense of computer vision, therefore the features learned can be applied to other tasks [59]. Thus, several well established CNNs such as AlexNet, GoogLe-Net, ResNet, VGG16, and others have been pre-trained on the ImageNet dataset and successfully used in a wide variety of computer vision tasks including detection, segmentation, and classification of medical images [61, 63, 64]. These pre-trained CNNs can be used as a feature extractor in which the top fully connected layers can be removed, and the output feature map can be flattened into a feature vector that can be used to train another separate machine learning classifier for different medical imaging application tasks [66].

Using these pre-trained CNNs leads to a third problem; there are many fundamental differences between the natural images in the ImageNet dataset used to train the CNNs and medical images [62]. The ImageNet images are natural color images with three channels (RGB), while mammograms are single channel greyscale images. Since the CNNs are trained on three channel RGB images they require this as an input. Stacking the grayscale mammogram into three channels is the most obvious solution, but this may provide redundant information to the network. While many studies have demonstrated potential or success in using transfer learning with the ImageNet dataset for classification of medical images [108-110], more work must be done to explore the role of image pre-processing and deep transfer learning for breast lesion classification using mammograms.

Although over the last decade great research efforts have been made to develop novel traditional and deep learning CAD schemes for detection and diagnosis of diseases [10, 110-112], the existing CAD schemes have their unique characteristics including different advantages and limitations, which have not yet been fully investigated or compared in previous studies [77, 78]. In our research work, we hypothesize that the traditional handcrafted features and the deep learning model generated features contain complementary discriminatory information because some of the handcrafted features can closely mimic image features or markers used by radiologists in lesion diagnosis, while automated features have the potential to extract new clinically relevant features that may be invisible or difficult to detect by human eyes. Thus, optimal fusion of these two types of features has potential to increase CAD performance to classify breast lesions. To test this hypothesis, this work aims to develop and evaluate a new fusion CAD scheme with improved mammogram lesion classification performance by combining both handcrafted and DL image features. The rest of the paper is organized as follows. Section 2 describes the information of the image dataset and experimental design including all steps of the proposed fusion CAD scheme. Section 3 reports study results by comparing lesion classification performance of CAD schemes using several machine learning classifiers trained using different sets of features. Section 4 discusses the impact of this study along with limitations and future work. Section 5 concludes this study.

3.2. Methods

3.2.1. Image Dataset

In our research laboratory, we have assembled a retrospective breast cancer screening image database of full-field digital mammograms (FFDM) under an institutional review board (IRB) approved image collection protocol. Each collected study case contains sequential FFDM images acquired in two to six annual screening sessions from 2008 to 2014. All FFDM images were acquired using the Hologic Selenia digital mammography machine (Hologic Inc., Bedford, MA, USA) with a fixed pixel size or spatial resolution of 70μm. Since in developing CAD schemes of mammograms, the high resolution FFDM images are used to detect microcalcifications and subsampled low-resolution images are used to detect soft tissue mass lesions, the original FFDM images are also subsampled using a pixel value averaging method with a 5×5-pixel frame to make image size of 818×666 pixels. Then, the subsampled FFDM image has a pixel size or spatial resolution of 0.35mm. The 12-bit gray level remains the same. From this image database, we have selected and assembled many subsets of images for different CAD tasks including predicting cancer risk, detecting and classifying suspicious breast lesions as reported in our previous research papers[35, 104, 113-115].

In this study, we collected 1,535 craniocaudal (CC) FFDM images from our existing image database. Each image contains a suspicious soft tissue mass-based lesion that was previously detected by a radiologist during the original image reading and diagnosis in screening environment. All lesions were recommended for biopsy. Based on the pathology examination results of the biopsied lesion samples, 740 lesions were confirmed as malignant, while 795 lesions were confirmed as benign. In each image, the lesion center was marked by the radiologist and recorded in our database. Using the recorded lesion center as a center reference, a square patch or region of interest (ROI) with a size of 64×64 pixels is extracted from each image. **Figure 3-1** shows two FFDM images (one malignant and one benign) and the corresponding ROIs.

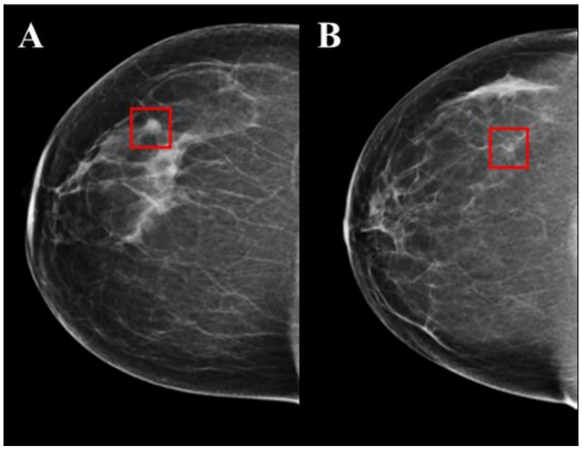

Figure 3-1: *A and B display two craniocaudal mammogram images including a malignant and benign lesion, respectively. Red boxes represent the 64x64 patch extracted around the suspicious lesion*.

3.2.2. Image preprocessing

In order to use deep transfer learning method, we first expand or rescale the original ROI with 64×64 pixels to 224×224 pixels using bilinear interpolation method. In order to generate three channel images suitable for deep transfer learning, we combine the original greyscale image (**I₀**) with two preprocessed variations [116]. Since mammograms are low dose X-ray images, these images may have poor contrast and the brightness may vary greatly between patients [117]. Firstly, a histogram equalization technique is applied to the original greyscale image to normalize and enhance the contrast of the mammogram (**$\mathbf{I}_{\mathbf{HE}}$**). Second, to denoise the mammogram images, we apply a bilateral low-pass filter to the original greyscale image (**$\mathbf{I}_{\mathbf{0}}$**) and generate a new filtered image (**$\mathbf{I}_{\mathbf{BF}}$**). This filter is selected because of its ability to reduce image noise, while effectively preserving edge and other textural information [118].

$$I_{BF}(p) = \frac{1}{W_p} \sum_{q \in \Omega} I_0(q) G_{\sigma_s}(||p - q||)\, G_{\sigma_r}(I_0(p) - I_0(q)) \tag{3-1}$$

where G_{σ_s} and G_{σ_r} are two Gaussian functions with two different kernel sizes determined by two sigma values, σ_s and σ_r, respectively. W_p is a normalization factor:

$$W_p = \sum_{q \in \Omega} G_{\sigma_s}(||p - q||)\, G_{\sigma_r}(I_0(p) - I_0(q)) \tag{3-2}$$

Thus, in using the bilateral low-pass filter, the first Gaussian low-pass filter (G_{σ_s}) in the spatial domain ensures that only the pixels in the area around the central pixel are considered and blurred. The second Gaussian low-pass filter (G_{σ_r}) considers the difference in intensity between pixels, which decreases the influence of pixel blurring with the increase of intensity difference and allows for edge preservation as edge locations have large intensity variations. As a result, applying this bilateral filter to mammograms will ensure that only pixels with small intensity variations (i.e., relatively homogeneous breast tissue or internal tumor areas) are blurred to reduce image noise, while tumor edge and other textural information of tumor and surrounding fibro-glandular tissues are preserved. Based on a previous study, the diameter of the pixel neighborhood (Gaussian filter kernel) used in this study is set to 9 and both sigma values (σ_s and σ_r) are set to 75 [116]. Therefore, three images, $\mathbf{I_o}$, $\mathbf{I_{HE}}$, $\mathbf{I_{BF}}$ are stacked to form a pseudo color image (**Figure 3-2**), which are fed to the CNN for automated feature extraction.

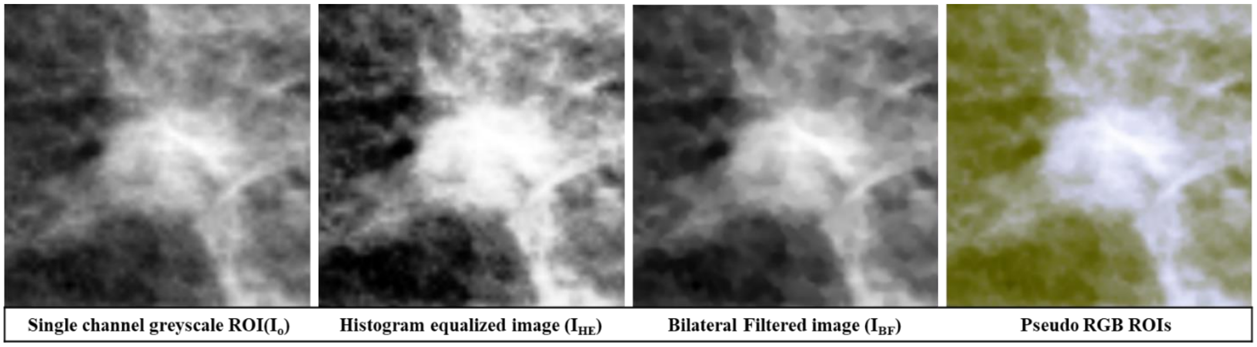
Single channel greyscale ROI(I_o)
Histogram equalized image (I_{HE})
Bilateral Filtered image (I_{BF})
Pseudo RGB ROIs

Figure 3-2: *Intermediate images in the creation of the pseudo-ROIs. Pseudo-ROI is created by stacking the three greyscale images.*

3.2.3. Deep Transfer Learning Feature Selection

Although several deep learning models have been applied as feature extractors for transfer learning in the medical imaging field, we use a VGG16 network whose weights have been pre-trained on the ImageNet dataset [119] as this network has performed well in many previous studies [120-123]. The VGG16 network is comprised of 13 convolutional layers followed by two full connected layers and a SoftMax layer [124]. All convolutional layers use a 3×3 kernel and all max-pooling layers have a stride of 2 (**Table 3-1**). VGG16 takes a 224×224 3-channel RGB image as an input. For the purpose of this study, the top fully connected layers are removed and a 7×7×512 feature map is extracted after the final max pooling layer. This feature map is then flattened into a 25,088-dimensional feature vector, which can be used to train a machine learning classifier.

Block	Layer	Size	Filter Size
1	Convolution-1	224×224×64	3×3
	Convolution-2	224×224×64	3×3
	Max pooling	112×112×64	-
2	Convolution-1	112×112×128	3×3
	Convolution-2	112×112×128	3×3
	Max pooling	56×56×128	-
3	Convolution-1	56×56×256	3×3
	Convolution-2	56×56×256	3×3
	Convolution-3	56×56×256	3×3
	Max pooling	28×28×256	-
4	Convolution-1	28×28×512	3×3
	Convolution-2	28×28×512	3×3
	Convolution-3	28×28×512	3×3
	Max pooling	14×14×512	-
5	Convolution-1	14×14×512	3×3
	Convolution-2	14×14×512	3×3
	Convolution-3	14×14×512	3×3
	Max pooling	7×7×512	-
6	Flatten	25,088	
	Dense	4,096	
	Dense	4,096	
	Dense	1,000	

Table 3-1: *VGG-16 Architecture. For this study, block 6 is removed and features are extracted after the final max pooling layer.*

Due to the extremely high dimensionality of this feature vector, a three-step feature reduction pipeline is used to select the optimal feature set. Since the VGG16 network is pre-trained on the ImageNet dataset that comprises a very heterogenous set of natural images, yet the intensity distributions of mammograms are relatively homogenous, a large percentage of the neurons will not be activated when a ROI of mammogram passes through. This would result in many features being inactive (zero) for most images. Thus, in the first step, all features with a variance of 0.01 or less are eliminated.

The second step in the feature selection pipeline takes advantage of a quick and powerful relief-based algorithm, Relief-F, to rank feature importance based on how well that feature does at differentiating between instances that are nearby [125]. It relies on a nearest neighbor approach to do so. Briefly, given a randomly selected instance R_i, represented by an a-dimensional vector, where a is the total number of features, Relief-F searches for the k nearest neighbors from the same class, near hits H_j, and k nearest neighbors from a different class, near misses M_j [126]. Feature weights, W[A], are then updated according to equation 3-3:

$$W[A] = W[A] - \frac{1}{n * k}\left(\sum_{j=1}^{k} diff(A, R_i, H_j) - \sum_{j=1}^{k} diff(A, R_i, M_j)\right) \tag{3-3}$$

where n is the total number of training instances. The *diff* function (equation 3-4) is used to quantify the difference between the attribute at two nearby instances.

$$diff(A, R_i, I) = \frac{|value(A, R_i) - value(A, I)|}{\max(A) - \min(A)} \tag{3-4}$$

where I is either a nearby hit, H_j, or a nearby miss, M_j. Larger feature weights reflect features that will be more relevant for distinguishing between classes therefore more desirable. The weight of feature A, W[A], will be increased if randomly selected instance R_i and nearby instance I belong to the different classes and the feature values are different. W[A] will be decreased if randomly selected instance R_i and nearby instance I belong to the same class feature values are different. The update in the feature weight will be proportional to the difference between the feature values as seen in equation 4. While original proposals of relief based algorithms describe a relevance threshold value, τ, such that all features with W[A] > τ, will be selected as a relevant feature, it is often more practical to select a set number of features to be considered relevant [127, 128]. In this study, 10 neighbors were used, and the top 300 features were chosen to undergo further feature selection. A more in depth review of relief based algorithms can be found elsewhere [127]. The final step in the feature selection pipeline uses a sequential forward floating feature selector (SFFS) with 10-fold cross validation to select the final optimal feature set [129, 130].

As mentioned previously, VGG16 pretrained on ImageNet takes a 3-channel image as an input. Two methods are used to convert one channel grey level image to a three-channel image. Therefore, in addition to extracting features from the pseudo-color-ROI images described in section 3.2, we also extracted features from a second group of input images, namely, stacked-ROI images. The stacked-ROI images are created by stacking the same original greyscale ROI in three channels. Therefore, after using the VGG16 network as a feature extractor and reducing the feature set, two independent optimal vectors of DL features are created, namely, a pseudo-ROI feature vector and a stacked-ROI feature vector.

3.2.4. Handcrafted Feature Extraction and Selection

There exists an abundance of CAD schemes which use a wide variety of handcrafted features. We initially computed 40 commonly used features from two separate feature groups. The first group consists of the first order statistical features that describe the distribution of pixel intensities across the image. These include 6 features namely, the mean, maximum, standard deviation, energy, skewness, and kurtosis of pixel intensity values. While first order statistical features provide information about the intensity distribution of the image, they do not provide any insight into the relative spatial positions of these intensities. The second group consists of textural features which describe the spatial arrangement of the intensity distributions. These textural features include those derived from the gray level co-occurrence matrix (GLCM) and the gray level run-length matrix (GLRLM).

The GLCM describes the number of co-occurrences of two pairs of grey level intensities which are a specific distance apart [131]. From the GLCM, 6 features are computed along four angles, 0° ,45° ,90°, and 135°, and at a distance of one pixel, namely, contrast, dissimilarity, homogeneity, ASM, energy, and correlation. The mean and maximum values of these features are computed resulting in 12 GLCM features. The GLRLM describes the number of consecutive pixels that have a specific pixel intensity [132]. From the GLRLM, 11 features are computed along the same four angles, 0° ,45° ,90°, and 135°, namely: short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, run percentage, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, and long run high gray level emphasis. The mean and maximum of each of these features are computed resulting in 22 GLRLM features. Mathematical descriptions of these features are explained in detail elsewhere [43]. After these 40 features were initially extracted, a variance threshold of 0.01 is applied to remove irrelevant features (**Figure 3-3**). As a result, 17 features are selected to create an optimal vector of handcrafted features.

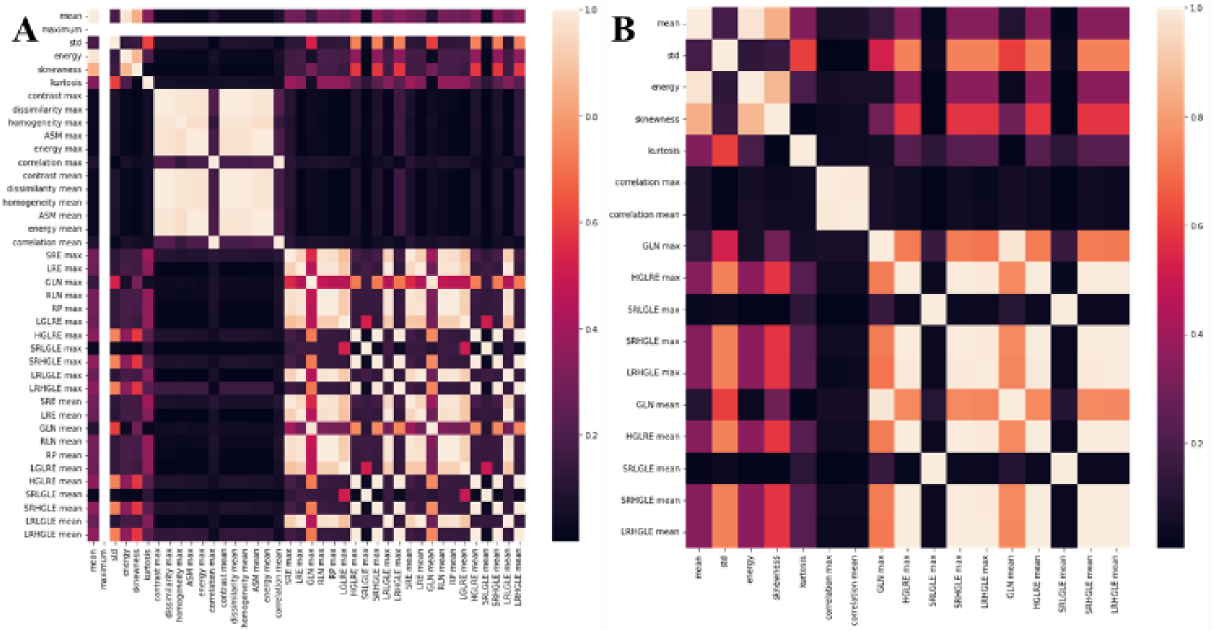

Figure 3-3: Heatmaps of correlation coefficients of the handcrafted features before and after applying variance thresholding.

3.2.5. Classification and Evaluation

Five separate machine learning classifiers are built using 5 optimal feature vectors extracted from handcrafted features, two sets of deep transfer learning features computed from pseudo-color-ROIs and stacked-ROIs, and fusion between handcrafted features and each set of automated features to test the hypothesis that fusion of handcrafted and deep transfer learning features can improve the performance of using CAD schemes to classify breast lesions as malignant or benign. Although many different types of machine learning classifiers have been used and tested in CAD field, we choose to use a support vector machine (SVM) as SVMs have many advantages as demonstrated in traditional machine learning or CAD field including its higher generalizability.

In this study, SVM1 is trained using only the handcrafted features, SVM2 is trained using the deep transfer learning features, which includes two SVM2s namely, SVM2-pseudo and SVM2-stacked, and SVM3 is trained using a fused feature set containing both the handcrafted and one set of deep transfer learning features, which also includes two SVM3s namely, SVM3-pseudo and SVM3-stacked. To build SVM3, handcrafted features and one set of deep transfer learning features are first combined through concatenation to create a new fusion feature pool. A SFFS algorithm is then used to select the optimal feature set from the fusion feature pool to train SVM3. All five SVMs are built using a linear kernel and trained using a 10-fold cross validation method. L2 regularization with C=1.0 is used to avoid overfitting.

Each trained SVM model is applied to every image in the testing fold and a prediction score between 0 and 1 is generated that indicates the likelihood of the image depicting a malignant lesion. Prediction scores of all 1,535 images are used to create a receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) is computed as an evaluation metric. Next, an operating threshold (T= 0.5) is applied on all SVM generated prediction scores to divide all testing images into two classes of malignant and benign lesions, so that the classification accuracy, inducing sensitivity and specificity, can be computed from a confusion matrix. In addition, the statistically significant differences of performance comparison are also computed and determined based on the criterion of $p<0.05$. A flowchart of this entire experimental design is shown in **Figure 3-4**.

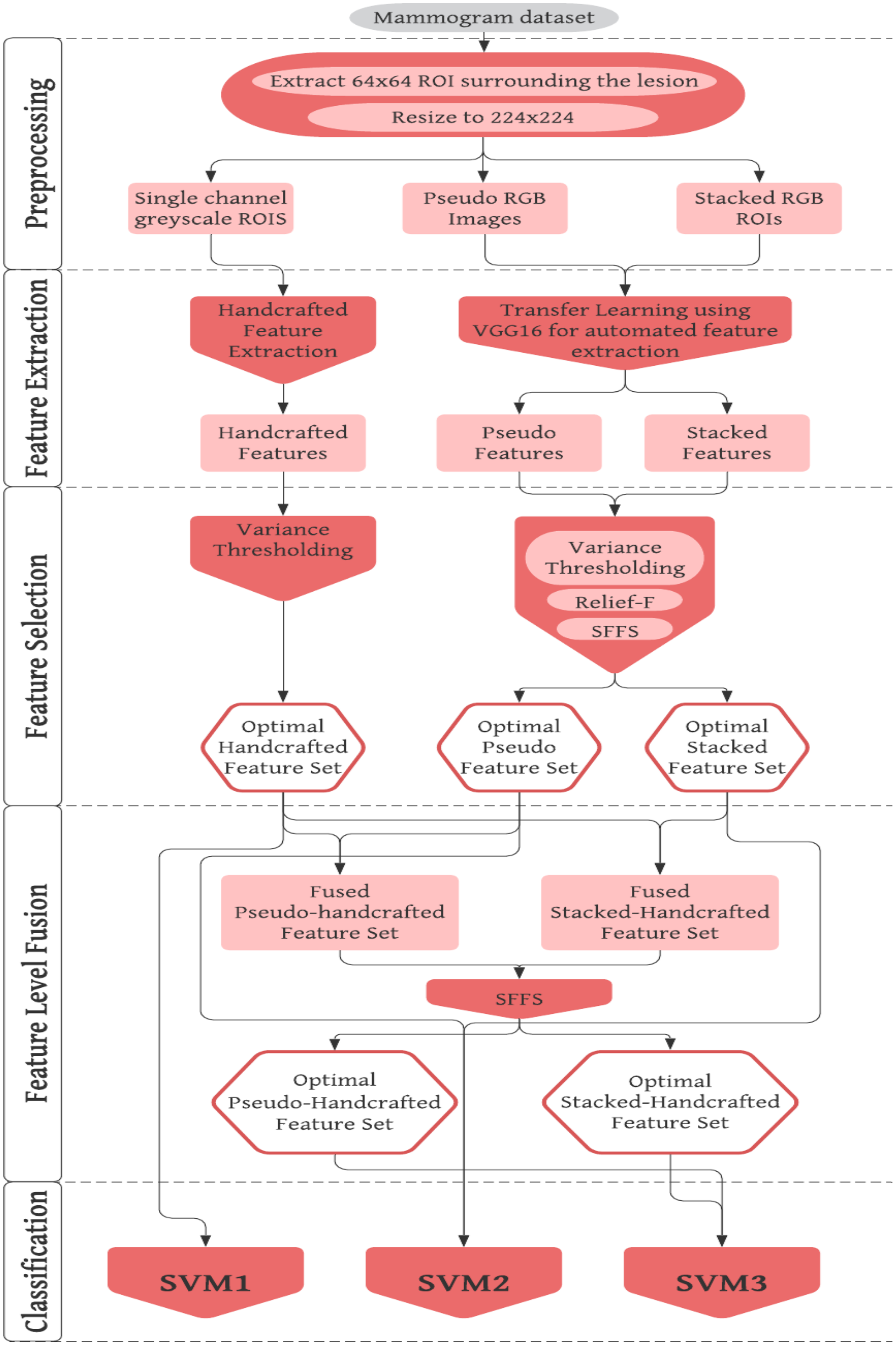
Mammogram dataset
Preprocessing
Extract 64x64 ROI surrounding the lesion
Resize to 224x224
Single channel greyscale ROIS
Pseudo RGB Images
Stacked RGB ROIs
Feature Extraction
Handcrafted Feature Extraction
Transfer Learning using VGG16 for automated feature extraction
Handcrafted Features
Pseudo Features
Stacked Features
Feature Selection
Variance Thresholding
Variance Thresholding
Relief-F
SFFS
Optimal Handcrafted Feature Set
Optimal Pseudo Feature Set
Optimal Stacked Feature Set
Feature Level Fusion
Fused Pseudo-handcrafted Feature Set
Fused Stacked-Handcrafted Feature Set
SFFS
Optimal Pseudo-Handcrafted Feature Set
Optimal Stacked-Handcrafted Feature Set
Classification
SVM1
SVM2
SVM3

Figure 3-4 *Flowchart of the entire experimental design.*

3.3. Results

Table 3-2 shows the results of the feature selection pipeline used to reduce the pseudo and stacked ROI feature sets. After variance thresholding about 75% of features were removed from the pseudo-ROI feature set and 70% were removed from the stacked-ROI feature set. This may support the idea that stacking the single channel greyscale image into a 3-channel RGB image provides redundant or irrelevant information. In addition, when applying a SFFS algorithm to select optimal features from two fusion pools of features, which include 17 handcrafted features plus 57 or 55 DL features (as shown in **Table 3-2**), two optimal fusion feature vectors including 61 and 37 features are generated. These two feature vectors include 9 and 6 handcrafted features, respectively.

Feature Selection Step	**Pseudo ROI**	**Stacked ROI**	**Handcrafted + Pseudo ROI**	**Handcrafted + Stacked ROI**
Initial number of features	25,088	25,088	74	72
Variance Thresholding	6,256	7,502	-	-
Relief-F	300	300	-	-
SFFS	57	55	61	37

Table 3-2 : *Number of features before and after feature reduction steps.*

Figure 3-5 shows 5 ROC curves generated from classification scores of 5 SVMs along with corresponding AUC values. When applying an operation threshold (T = 0.5), each ROI with an SVM-generated classification score ≥ T is classified as a malignant lesion, otherwise, it is classified as a benign lesion. **Figure 3-6** shows 5 confusion matrices generated by 5 SVMs. These confusion matrices represent the sum after 10-fold cross validation. Based on the data shown in **Figures 3-5 and 3-6**, the mean AUC values (along with corresponding standard deviation of 10-fold cross-validation), classification accuracy, sensitivity, and specificity of all five SVMs are summarized in table 3-3.

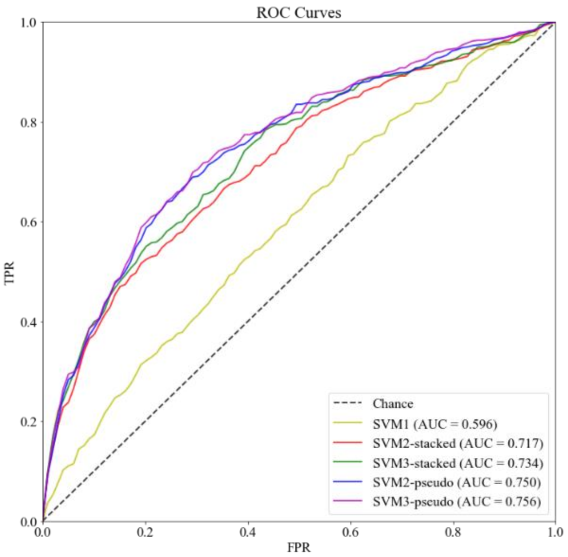

Figure 3-5: *ROC AUC curves for all 5 SVMs.*

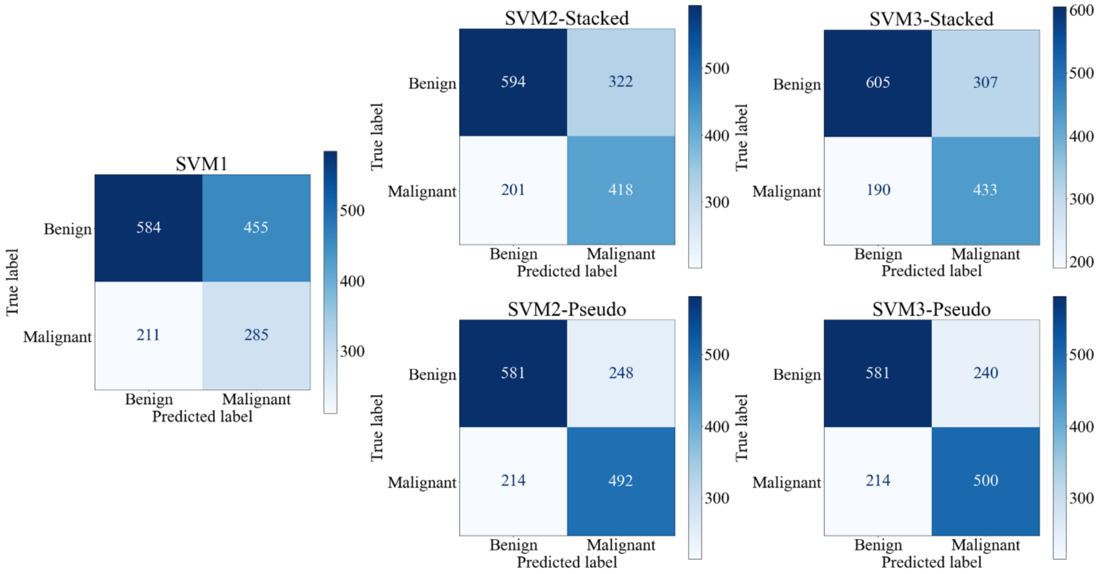

Figure 3-6: *Confusion Matrices of all 5 SVMS. Confusion matrices are the sum of all matrices after 10-fold cross validation*

Model	AUC	Accuracy	Sensitivity	Specificity
SVM1	0.596±0.032	0.566±0.033	0.385±0.055	0.735±0.044
SVM2-Stacked	0.717±0.022	0.659±0.030	0.565±0.052	0.747±0.037
SVM3-Stacked	0.734±0.017	0.676±0.041	0.585±0.058	0.761±0.032
SVM2-Pseudo	0.750±0.043	0.699±0.036	0.665±0.064	0.731±0.028
SVM3-Pseudo	0.756±0.042	0.704±0.035	0.676±0.061	0.731±0.028

Table 3-3: *Summary of classification performance indices including mean values and standard deviations of all 5 SVM models after 10-fold cross validation.*

A paired t-test at an alpha value of 0.05 was used to test for a statistically significant difference in means between classification performance of groups of two SVMs. When analyzing the three SVMs developed using features extracted from the stacked ROIs, we observe that SVM3-stacked, trained using a fused feature vector, performs significantly better than both SVM1, trained using handcrafted features, and SVM2-stacked, trained using DL features, in both AUC and accuracy (**Figure 3-7A and B**). When analyzing the three SVMs developed using features extracted from the pseudo-ROIs, we observe that the AUC value yielded by SVM3-psuedo is also significantly higher than AUC values yielded by both SVM1 and SVM2-psuedo (**Figure 3-7C**). While the accuracy of SVM3-psuedo is greater than that of SVM2-pseudo, this difference is not significant (p=0.1363) (Figure 3-7D). Since both feature fusion based SVMs performed better than both SVMs trained using single type of features, the study results validate and support our hypothesis that feature fusion by optimally selecting handcrafted and automated features can create a machine learning classifier with improved classification abilities.

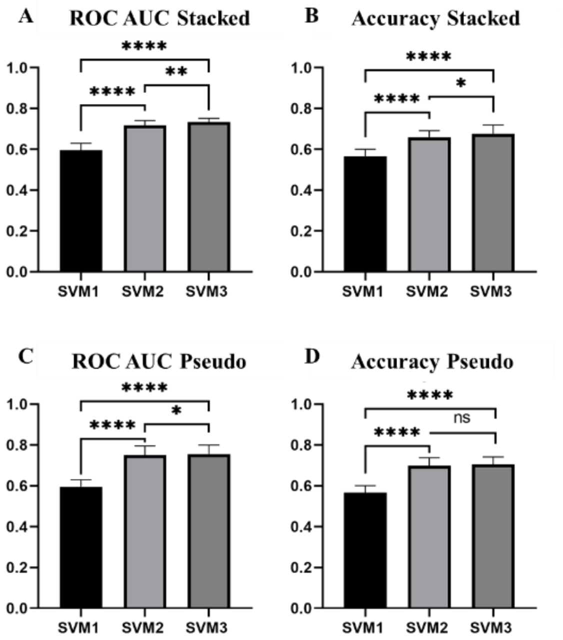

Figure 3-7: *Bar graphs displaying the mean and standard deviation (STD) of all three SVMs for the pseudo-ROIs and stacked ROIs. (**** = p <0.001, **=p<0.01, *=p<0.05)*

In addition to developing and evaluating a feature fusion-based SVM classifier for improved performance, we also compared the performance of the SVMs trained using pseudo-ROIs as an input to VGG16 for feature extraction, with the performance of SVM2s trained using stacked-ROIs as an input to VGG16 for feature extraction. While there is only a statistically significant difference between the accuracies of SVM2-stacked and SVM2-psuedo (p=0.0251), SVMs trained using features extracted from pseudo-ROIs achieve a higher AUC and accuracy than SVMs trained using features extracted from stacked-ROIs (**Figure 3-8**).

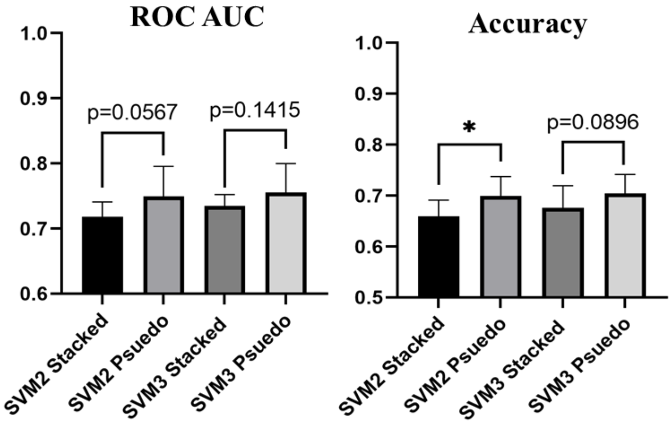

Figure 3-8: *Bar graphs displaying the difference in ROC AUC and Accuracy of SVMs trained using deep transfer learning features extracted based on pseudo-ROIs and stacked-ROIs (*=p<0.05)*

In addition, **Figure 3-9** shows 3 blocks or pairs of lesion classification examples. Each pair includes one malignant lesion (marked by red ROI frame) and one benign lesion (marked by yellow ROI frame) in the top row. The magnified images of the extracted ROI are shown in the bottom row of the figure. First, the two lesions in the left block are correctly classified by 3 SVMs (SVM1, SVM2-pseudo, and SVM3-pseudo). We can see from the ROIs that the benign mass has a roundish shape and looks relatively uniform when compared to the background, while the malignant lesion appears with spiculated margins and is much brighter than the surrounding tissue. Second, two lesions in the middle block are correctly classified by 2 SVMs (SVM2-pseudo and SVM3-pseudo), but incorrectly classified by SVM1. Since SVM1 is trained on handcrafted features only, this means that using only handcrafted features is not sufficient to make a correct distinction but DL features alone and the combination of the handcrafted and DL features can be more accurate. Third, the two lesions in the right block may be more subtle and are only correctly classified by one SVM (SVM3-pseudo) and misclassified by SVM1 and SVM2-

pseudo. This highlights the fact that handcrafted and DL features do contain complementary information that when used together can better classify suspicious lesions.

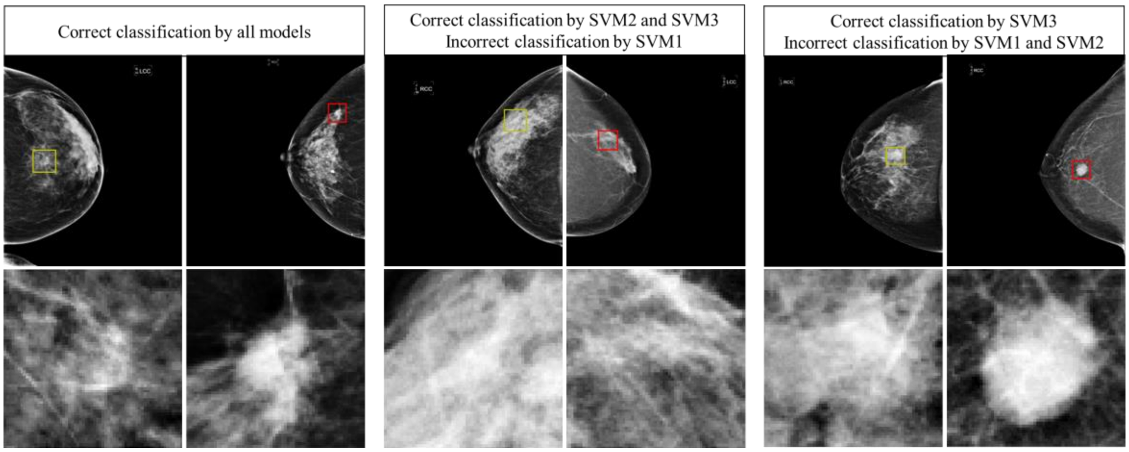
Correct classification by all models
Correct classification by SVM2 and SVM3
Incorrect classification by SVM1
Correct classification by SVM3
Incorrect classification by SVM1 and SVM2

Figure 3-9: *Examples of correct and failed classifications. The top row displays the full CC image while the bottom row shows the ROI that was used for feature extraction. Yellow ROI indicates that the true value is benign while a red ROI indicates that the true value is malignant.*

In summary, when using only deep learning features with pseudo-ROI input images, the AUC value and classification accuracy of SVM2-pseudo increase 25.8% (from 0.596 to 0.750) and 21.9% (from 0.566 to 0.690), respectively, as compared to SVM1 trained using handcrafted features only. Additionally, when fusion of handcrafted and DL features, AUC value and classification accuracy of SVM3-pseudo are further increased by 0.8% (from 0.750 to 0.756) and 2.1% (from 0.690 to 0.704), respectively, as comparing to SVM2-pseudo.

3.4. Discussion

This work demonstrates a new CAD scheme for the classification of breast lesions as malignant or benign. This study uses a diverse dataset of 1,535 cases, which is larger than most datasets used in previous CAD studies to classify breast lesions (such as 8 studies summarized in [133], which reported AUC values ranging from 0.70 to 0.87 using datasets involving 38 to 1,076 images). The reported performance of breast lesion classification in this study is not directly comparable to those reported by many previous studies due to the use of different image datasets. However, this new CAD scheme shows promising performance compared to the high rates of false-positive recalls and unnecessary biopsies of benign lesions in current clinical practice. The contribution of this study includes following unique characteristics or research approaches and new interesting observations, which fully support our study hypothesis.

First, many deep learning CAD schemes have been developed and reported in the literature. Previous approaches can be divided into three categories. (1) The studies that use deep learning as an end-to-end classifier. For example, using a smaller dataset of 560 FFDM images (280 are malignant and 280 are benign), Qiu et al. developed and tested a deep learning model using a 4-fold cross validation method [58]. The study reported AUC values ranged from 0.696±0.044 to 0.836±0.036 in 4 folds (with average of AUC = 0.790±0.019). (2) The studies that use a deep transfer learning model as feature extractors. For example, Mendel at al. used a pretrained VGG19 model to extract features from 78 biopsy confirmed FFDM cases [134]. Then, by using the extracted features to train a linear SVM classifier using a reduced feature set, the study reported an AUC of 0.76±0.05. This study is somewhat like the SVM2 classifiers trained in our study, which yielded comparable AUC=0.75±0.04 (SVM2-pseudo) while it is tested using a much larger image dataset. (3) The studies that fused two classifiers separately trained using handcrafted and automated features at the final decision level. For example, Huynh et al.

applied a soft-voting technique to fuse the outputs of two SVMs trained using automated features extracted by an AlexNet model and an SVM trained using traditional CAD features [80]. When applying 607 ROIs extracted from 219 FFDM cases and using a 5-fold cross-validation method, the study reported an AUC of 0.86±0.01. However, in this study, we investigate a new novel approach that fuses handcrafted and DL features at feature selection level to create an optimal feature set and train a single classifier (i.e., SVM3-pseudo or SVM3-stacked). To the best of our knowledge, such fusion method to develop CAD schemes of medical images has not been reported in the literature.

Second, since using a deep transfer learning model as a feature extractor generates a very large feature vector (25,088 from VGG16 model), identifying a small set of optimal features is a difficult but important task. Our experiments indicated that many commonly used feature dimension reduction methods including principal component analysis (PCA) have lower performance when applying to such large feature vectors. Thus, in this study, we developed a new feature selection pipeline with three steps that allows for the successful selection of an optimal set of automated deep learning features from the huge number of initially extracted features. Among these three steps, we investigated and identified an optimal and effective approach to use Relief based algorithms, which are unique in that they do not assume independence among features as many other feature selection filter methods do. Relief-F was chosen for this study since we are unaware of what kind of feature interactions exist from the feature map extracted from VGG16. A limitation of Relief-F worth noting is that in an extremely large feature space the identification of a nearest neighbor becomes increasingly random, which leads to a decrease in performance [126, 135]. Iterative RBA such as Iterative Relief [136], Tuned

Relief-F(TuRF) [137] , VLSRelief-F [135], and more [138], have been developed which improve the performance in very large feature spaces. There is no consensus on what defines an extremely large feature set or when these iterative approaches perform better. We observed no significant difference in the performance of SVM2 or SVM3 when using Relief-F alone and using TuRF wrapped around Relief-F. As a result, the optimal Relief-F algorithm was used to reduce dimensionality of feature space by more than 95% (i.e., reducing the number of features from 6,256 to 300 when using pseudo-ROIs). Combining the three steps in this feature selection pipeline, the number of features is reduced to 55 or 57 from original 25,088 (as shown in Table 3-2), which supports building robust SVMs using a large training dataset with 1,382 cases (9 folds of our dataset).

Third, this study supports that using deep transfer learning model generated features has significant advantages over using the traditional handcrafted radiomics features since classification performance of SVM2 is significantly higher than SVM1 (i.e., AUC=0.750 for SVM2-pseudo and AUC=0.596 for SVM1). However, our study also demonstrates that the handcrafted features and DL features contain complementary information to classify breast lesions. Thus, using the fusion feature sets including both handcrafted and DL features to train and build SVM3-pseudo and SVM3-stacked enables to further improve lesion classification performance. Since both handcrafted and deep features extracted from mammograms may be able to pick up on details and patterns that cannot be detected with the human eye, classifications made by this fusion-based CAD scheme have the potential to better assist radiologists in reducing the false positive recall rate of mammogram lesion detection by acting as a second reader.

Fourth, we observe that AUC value and classification accuracy are higher when using pseudo-ROIs for feature extraction from deep transfer learning models than simply using 3 stacked ROIs. Few studies have been conducted to investigate how to optimally utilize pseudo-color ROIs as inputs for deep CNNs. Overall, these studies show higher lesion classification performance when using ROIs that have been meaningfully pre-processed when the original greyscale image is just stacked in three dimensions [116, 139, 140]. This further solidifies the idea that image preprocessing is a crucial step when utilizing a deep learning network trained on natural images for medical imaging tasks. As many different contrast enhancement techniques exist for processing mammograms [117], future work must be done to better investigate these techniques roles in developing more effective pseudo-RGB images for deep transfer learning.

Despite a large image dataset, promising classification results, and new observations that can help facilitate research effort to further develop and optimize CAD schemes to classify between breast lesions using mammograms, there are also several limitations in this study. First, the dataset used in this study focuses solely on craniocaudal view mammograms. As a result, this is only a region-based CAD scheme. Since in mammography a lesion can often be detected in both craniocaudal (CC) and mediolateral oblique (MLO) views, fusion of classification results from two views has potential to further improve classification performance. Thus, in future studies, we will develop and test a more complete case or lesion-based CAD scheme that fuses the two lesion regions detected on both CC and MLO views. Second, for a proof-of-concept study we only computed 40 handcrafted image features, used VGG16 as a deep transfer learning model as a feature extractor, and a standard SVM as a multifeatured-based classifier. Although this is an efficient approach to test our hypothesis, the results may not be the best or optimal. More existing radiomics features and/or other features (i.e., local binary patterns) should be explored in handcrafted features, and more advanced deep transfer learning and classification models should be investigated and applied in future studies to improve lesion classification performance. Third, as shown in figure 3-4, this study used a concatenation method as a feature-level fusion operator. Although the concatenation method is widely used in the CAD field, it may not be the best method. We will further investigate different feature-level and decision-level fusion schemes in future studies. Fourth, we recognize the importance of using balanced dataset of two classes to train machine learning classifier. In this study, our dataset is slightly unbalanced with a ratio of 1.074 (795/740 or 51.8% benign and 48.2% malignant images). Although the imbalanced ratio is relatively small and should not have significant negative impact in training SVM models, we will investigate this issue or impact of using the more balanced image datasets such as adding synthetic FFDM images using a Generative Adversarial Networks (GAN) as demonstrated in a recent study in our lab [141]. Lastly, although we use a standard 10-fold cross-validation method to test CAD performance, its robustness needs to be further validated or tested on multiple different datasets and compared to different traditional or deep learning classifiers in future work.

3.5. Conclusions

This study presents a new fusion-based CAD scheme that combines handcrafted features with automated deep transfer learning features aiming to improve the performance of a machine learning classifier in classification of breast lesions as malignant or benign. Although this is a preliminary study with several limitations, to the best of our knowledge, this is the first proof-of-concept study that investigates and demonstrates the feasibility and advantages of optimally fusing handcrafted and two types of deep transfer learning generated automated features extracted from pseudo-ROIs and stacked ROIs to train new machine leaning classifiers to improve accuracy in breast lesion classification. Therefore, this study helps build a solid foundation for us to facilitate future studies and make progress in this CAD field. We currently continue to investigate new approaches to (1) compute both handcrafted features (based on radiomics concept) and automated features (based on improved deep transfer learning models) including using more effective image pre-processing method to produce pseudo images, (2) more effectively post-process the automated features to generate optimal and more robust feature vectors to train machine learning classifiers, and (3) to investigate and apply more effective fusion methods to combine handcrafted and automated features to train machine learning classifiers, which aims to more effectively take or combine advantages of both types of image features.

Chapter 4. A multi-stage fusion framework to classify breast lesions using deep learning and radiomics features computed from four-view mammograms

4.1. Introduction

Early detection of breast cancer is critical for improving the efficacy of cancer treatment and patient survival. For the last several decades, routine population based mammography screening has played a crucial role in early cancer detection and is one of the primary reasons for the decrease in the mortality rate of breast cancer.[2] Despite the widespread utility of mammography or digital breast tomosynthesis (DBT) recently, the efficacy of these population-based breast cancer screening exams is low and controversial due to the high false-positive recall and benign biopsy rates.[142] As a result, decreasing the false-positive rates is a pressing clinical issue as it leads to unnecessary biopsies which often have long-term psychological consequences on the patients in additional to being an economic burden on the society.[143]

Thus, to help radiologists more accurately detect suspicious breast lesions and distinguish between malignant and benign lesions, computer-aided detection (CADe) and diagnosis (CADx) schemes have been extensively developed over the last several decades. Currently, commercialized CADe schemes have been used in the clinical practice to assist radiologists in detecting suspicious lesions while reading mammograms.[10] However, whether using CADe can add real clinical values remains questionable[76] due to the higher number of false-positive detections.[3, 4] On the other hand, CADx schemes which have the goal of helping radiologists more accurately classify between malignant and benign breast lesions detected on the mammograms to reduce false-positive recall rate and the number benign biopsies have not yet been accepted or applied in clinical practice to date. Difficulties with current mammogram-based CAD systems (both CADe and CADx schemes) arise from (a) low contrast images intrinsic to X-ray mammography that require various pre-processing methods, (b) drastic differences in the spatial location and appearance of suspicious lesions (i.e., soft tissue masses) which make it difficult to obtain a large and diverse training dataset, and (c) the plethora of breast segmentation schemes with no consensus on the best method to use. Therefore, there is still a need to improve the performance of mammography-based CAD systems and the manner in which these systems are employed.[3, 4]

In a typical mammography screening exam, two projection images are taken of each breast namely, a craniocaudal (CC) view and a mediolateral oblique (MLO) view, resulting in four images per screening exam (left-CC (LCC), right-CC (RCC), left-MLO (LMLO), and right-MLO (RMLO)) (**Figure 4-1**). When a radiologist reads mammograms from one screening exam, he/she combines information from all four view images to decide if a suspicious lesion is present or not and whether the presented lesion is malignant or benign (or whether the patient should be recalled for an additional exam or biopsy). However, most existing CAD schemes are either single image-based (CADe) or lesion-based (CADx) schemes in which the CAD schemes only analyze information or image features extracted from a single view image. This is thought to be one major reason that limits the performance of current CAD schemes, particularly, its capability to reduce false-positive detections (for CADe schemes) and classify lesions (for CADx schemes). As a result, this has led to an increase in research focused on exploring new technologies and approaches to effectively compute matched multi-view information or image features from multiple mammograms and how to optimally fuse the multi-view image features to build new multi-view image-based CAD schemes.[3, 4, 88]

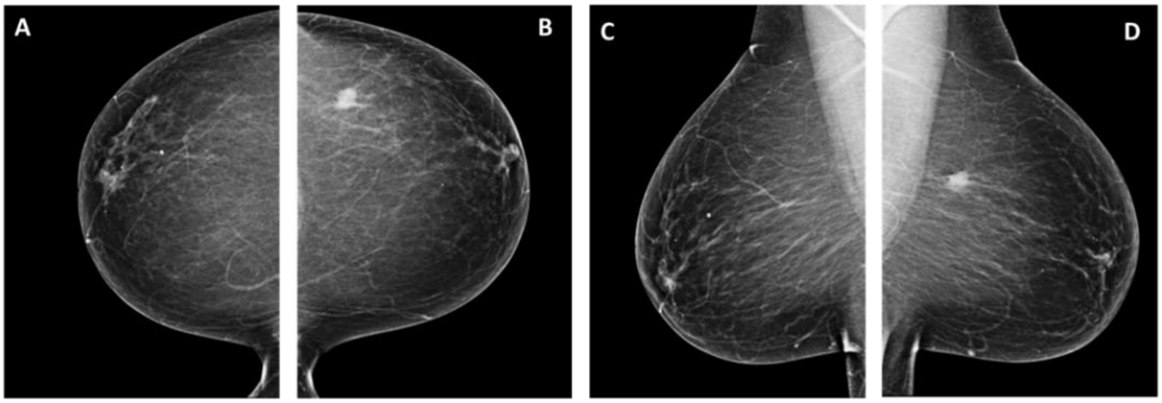

***Figure 4-1**: Example of the CC and MLO projection views taken in mammography, which are named as (A) RCC, (B) LCC, (C) RMLO, and (D) LMLO images, respectively.*

Although approaches to develop multi-view CAD schemes have been proposed and reported in the literature, they can be divided into three categories. The first and most popular method uses ipsilateral views by fusing information from CC and MLO views of one breast, which allows for the extraction of image characteristics that may be obscured due to dense overlapping fibro-glandular tissue in one projection view but visible in the other projection view. The second method uses bilateral views by fusing information from the same projection view of the left and right breast, which allows for quantification of breast tissue asymmetry (i.e., parenchymal distortions or change in contrast). This method mimics how radiologists tend to pay careful attention to the asymmetry between bilateral breasts as highly asymmetrical breasts is often a good indicator of breast cancer and the locations of asymmetry often contain suspicious lesions (i.e., masses).[144] The third method uses both the ipsilateral and bilateral views by fusing information from all four images, which aims to take advantages of methods one and two.

Developing multi-view image-based CAD schemes often faces several challenges including a difficult image registration task. Mammography exams require the breast to be compressed, as a result simple rigid or affine transformation techniques cannot properly model the anatomical deformations present in the compressed breast. One technique commonly used to accomplish this non-rigid registration task is to use a free-form deformation (FFD) field parametrized by a B-spline control point mesh.[81] However, many studies bypass this difficult registration task by using basic subtraction techniques without image registration and ignoring the mismatch between breast regions which results in inaccurate asymmetry quantifications.[82, 145] Other studies do not quantify the bilateral asymmetry at all and just use whole breast images of bilateral breasts independently. Despite these difficulties, previous studies have shown that regardless of using either two view images (from only ipsilateral or only bilateral views) or a combination of four-view images, multi-view CAD schemes consistently outperform single-view CAD schemes, indicating that the information contained in different projection views of bilateral breasts can provide additional useful information in detecting and classifying suspicious breast lesions from mammograms.[84]

The jump from single-view to multi-view CAD schemes introduces another issue as developers must also consider how to efficiently extract and fuse information from multiple input images. Traditionally, a set of handcrafted radiomics features would be extracted from the input image and then used to train a machine learning classifier. More recently, deep learning models are used to extract information directly from the input image based on learned representations of a target domain. While deep learning-based CAD schemes tend to outperform conventional machine learning based CAD schemes, they require extremely large amounts of input data for adequate training and testing, which is often not available in the medical imaging domain. Hence, handcrafted radiomics feature extraction is still a relevant technique. Additionally, handcrafting specific radiomics features benefits from prior knowledge of the domain, meaning image characteristics that are known to be relevant to the task can be quantified through mathematical models and used as image features. On the other hand, deep learning-based features thrive in areas that traditional features lack since deep learning-based features can capture patterns that may not be distinguishable by human eyes, therefore, cannot be quantified by a human crafted mathematical model. Several studies have investigated potential advantages of combining handcrafted radiomics features with automated deep learning-based features to improve model classification performance.[89-91] However, there is no consensus on the best way to fuse the information extracted from multiple input images using multiple feature extraction methods.

As outlined above, the three main considerations when developing CAD of mammograms are (1) single-view or multi-view schemes, (2) multi-view schemes based on ipsilateral-view analysis or bilateral-view analysis or both, and (3) the schemes using traditional radiomics features or deep learning generated features. Previous studies demonstrate that multi-view CAD models tend to outperform single-view CAD models,[88] the addition of information from the contralateral breast to quantify the bilateral asymmetry increases model performance[89-91], and the fusion of handcrafted radiomic features and deep learning features outperforms either method alone when classifying suspicious breast lesions[41, 79]. However, to the best of our knowledge, no existing work combines these three points into a singular framework. In order to address these challenges, we hypothesize that (1) the automated features generated by deep transfer learning model and handcrafted radiomics features contain complementary information, and optimal fusion of these two types of features can improve CAD performance in tumor classification, and (2) a 4-view image-based CAD scheme can yield significantly higher tumor classification performance than one or two-view image-based CAD schemes. In the rest of this paper, CAD scheme refers to CADx scheme of lesion diagnosis or classification. To test our hypothesis, this study systematically investigates and compares advantages and limitations of fusing deep learning generated features and traditional radiomics based features to develop multi-view CAD schemes.

Specifically, this study focuses on the investigation of the following issues, namely, (1) identifying and extracting matched regions of interest (ROIs) from four mammograms, (2) exploring and computing a new type of image features to represent bilateral image feature asymmetry, and (3) training and testing machine learning classifier using different image feature fusion methods namely, feature level and output level fusion techniques. Through these investigations, the goal of this study is to demonstrate the feasibility of developing a new optimal case-based CAD framework to classify suspicious breast lesions, which fuses both handcrafted radiomics (HCR) features and deep transfer learning (DTL) features computed from four CC and MLO view mammograms of the left and right breasts. Detailed descriptions of the technical development of this framework are presented in the following sections.

4.2. Materials and Methods

4.2.1. Image Dataset

The dataset used in this study is assembled from a large de-identified image database of full-field digital mammograms (FFDM). These FFDM images were acquired under an institutional review board approved image collection protocol using the Hologic Selenia digital mammography machine (Hologic Inc., Bedford, MA, USA) from 2008 to 2014. Detailed image and dataset characteristics can be found in our previous studies.[145] In brief, sizes of the original FFDM images are either 3,325×4,095 or 2,555×3,325 pixels with one-dimensional pixel size of 0.07mm. To develop CAD schemes of mass-type lesion detection or classification, an average kernel with 5×5 pixels is applied to subsample each original FFDM image. As a result, sizes of FFDM images are reduced to 665×819 or 511×665 pixels with pixel size of 0.35mm. In this study, we selected all available cases that have four FFDM images representing CC and MLO view of the left and right breast. Each case contains one suspicious mass-based lesion that has been marked by a radiologist and proven by biopsy as malignant or benign. Cases were excluded if the mass was not visualized in both CC and MLO view.

To confirm that the mass seen on both CC and MLO view is the same mass, an ipsilateral matching process was applied. Prior to ipsilateral matching, background artifacts are removed by first converting the image to a binary image using an Otsu thresholding method and then creating a mask based on the external contour.[146] The mask is applied to the original image, resulting in an image of the whole breast region with all background artifacts removed. The first step of the ipsilateral matching process is to identify the location of the pectoral muscle in both views. The pectoral muscle is often not visible in the CC view; therefore, the location of the pectoral muscle on the CC view is defined as a vertical line that is parallel to the edge of the image. To identify the pectoral muscle on MLO images, a straight line approximation is made based on the average gradient as adopted from a previous study.[147] The pectoral muscle location is then used to identify the nipple location in both CC and MLO images following the method developed in previous study.[148]

Once these landmarks are identified, ipsilateral matching is conducted based on an existing method.[149] Briefly, the centerline is first defined which is a line perpendicular to the pectoral muscle that passes through the nipple. Next, the mass is projected onto the centerline and the distance between the mass projected onto the centerline and the nipple is calculated (**Figure 4-2**). If the absolute difference between this distance from the CC view and MLO view is less than 100 pixels, then the two masses are considered ipsilaterally matched. In the subsampled images, 100 pixels represents 35mm which should comfortably match small and large lesions while accounting for bias introduced by the radiologist when marking the center of each lesion.[150] Masses that could not be matched ipsilaterally are discarded.

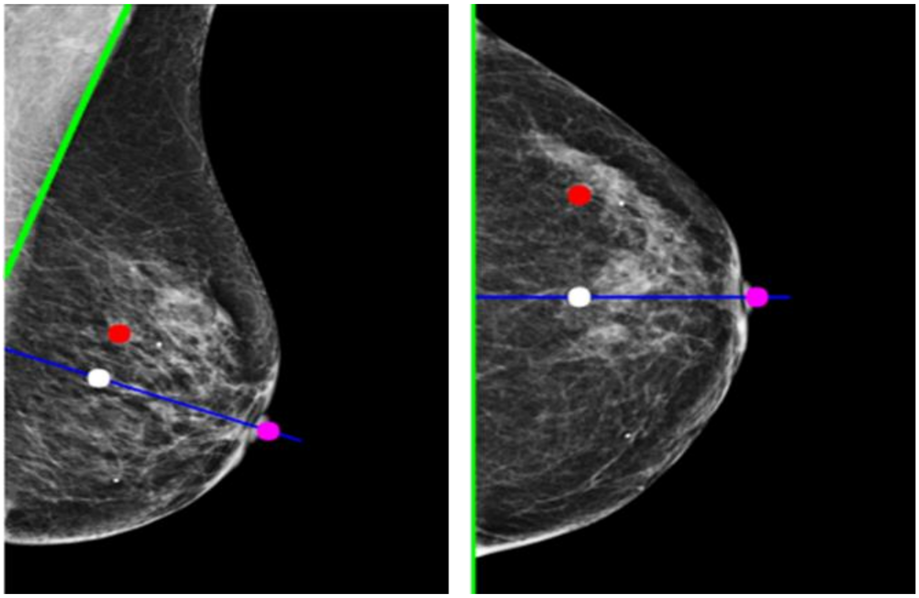

***Figure 4-2**: Example of the ipsilateral matching scheme. The location of the pectoral muscle is drawn in green, the location of the nipple is shown by a pink dot, and the centerline is drawn in blue. The LMLO and LCC images in this case each contained one suspicious lesion as marked with a red circle. After the centerline is drawn, the mass is projected onto the centerline (white dot) and the distance to the nipple is calculated. For this case, the distance was 157.93 pixels for the LMLO view and 155.00 pixels for the LCC view. Since the absolute difference between these values is less than 100, we consider this mass to be ipsilaterally matched.*

4.2.2. A new multi-view CAD framework

Figure 4-3 is a visual representation of the proposed multi-view CAD framework developed and tested in this study. In this figure, we assume that a suspicious lesion is visually detected on FFDM images of the right breast. Thus, two suspicious lesion regions (ROIs) are located and extracted on both CC and MLO view images of right breast, which are defined as RCC image and RMLO image on the top row of the figure. Then, multiple CAD image processing and feature analysis steps are applied to build machine learning classifiers to predict the likelihood of the queried suspicious lesion being malignant. The details of all CAD steps are described below.

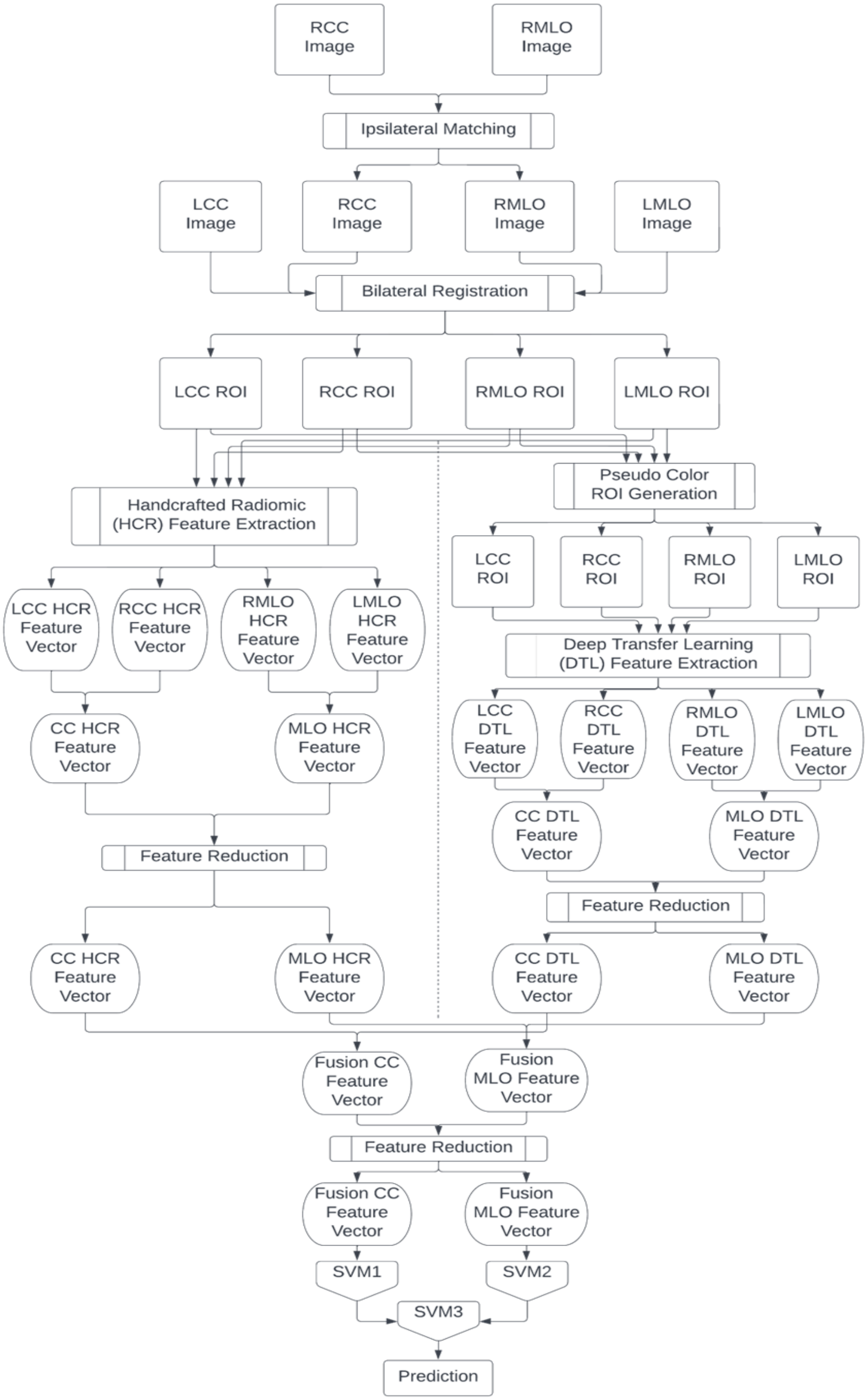

Figure 4-3: Flowchart of the framework of this study

4.2.2.1 Extraction of matching regions in four views

As shown in **Figure 4-3**, after ipsilateral matching, all cases are represented by four images where two of those images are ipsilateral views of the same suspicious lesion and the other two images are ipsilateral views of the contralateral breast without a lesion. To quantify the bilateral asymmetry of image features computed between bilateral images, we perform bilateral image registration to identify and extract two matched regions of interest (ROIs) from two bilateral mammograms, which includes two pairs of the registered ROIs (LCC-RCC and LMLO-RMLO).

Before registration, all right breast images are mirrored so that the orientation of the left and right breasts are the same. Bilateral registration is conducted using a multiresolution B-spline transformation that optimizes the mattes mutual information metric.[151] The registration method is implemented using SimpleITK of the Insight Toolkit (ITK) in python.[152] For this study, the mammogram containing a suspicious lesion is used as the fixed image while the contralateral mammogram is used as the moving image. Registration is conducted in this manner so that the annotations of the center of the suspicious lesions remain accurate. Once the images are bilaterally registered, four matched ROIs of 64×64 pixels are extracted surrounding the center of each lesion region on two ipsilateral view of the same breast and from two matching ROIs on two ipsilateral views of the contralateral breast (**Figure 4-4**).

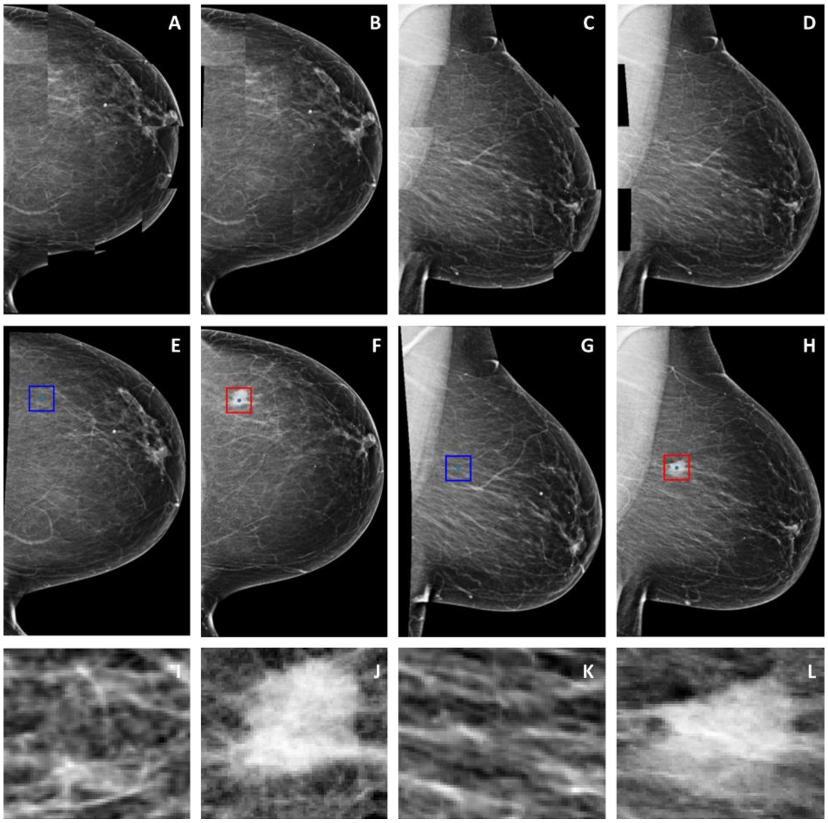

Figure 4-4: *Example of bilateral registration and ROI extraction. The top row displays the B-spline transformation via checkerboard visualization. (A) is the unregistered CC images, (B) is the registered CC images, (C) is the unregistered MLO images, and (D) is the registered MLO images. The middle row displays the registered bilateral images for the (E) RCC, (F) LCC, (G) RMLO and (H) LMLO view. The red bounding boxes show the 64x64 pixel ROI extracted surrounding the center of the lesion as marked by a radiologist. Blue boxes seen in the contralateral images show the location in which the corresponding ROI is extracted after bilateral registration. The bottom row shows the extracted ROIs from the corresponding view above it.*

4.2.2.2 Handcrafted Radiomics Feature Extraction and Reduction

Forty-five handcrafted radiomics (HCR) features are first computed from each ROI independently. These include 7 first-order statistical features that describe the intensity distributions of the images with no spatial information, 12 gray-level cooccurrence matrix (GLCM) derived features and 22 gray-level run length matrix (GLRLM) derived features that are used to describe the spatial distribution of the varying intensity distributions. Additionally, four Gabor features are extracted since these features are known to be extremely useful for mammography texture analysis as the filters have optimal Heisenberg joint resolution in the spatial frequency domain, so that the features are able to overcome the intrinsic low resolution and high noise of mammography images.[153] A Gabor filter bank of 16 filters is created from a combination of the following parameters, spatial frequency of the harmonic function of 0.05 or 0.25, orientation of 0-4, and standard deviation of the Gaussian kernel of 1 or 3. Each image is convolved with each filter and the mean, variance, energy, and entropy are calculated from the filtered image. Then, the mean of each feature over the 16 filtered images is taken resulting in four Gabor features per image.

After HCR feature extraction, four feature vectors are created namely, LCC-HCR, RCC-HCR, LMLO-HCR and RMLO-HCR, each containing 45 features. Next, each of these features $(f_i^{org}, i = 1,2,\cdots,45)$ is independently normalized using the following equation.

$$f_i^{Norm} = \frac{f_i^{org} - f_i^{min}}{f_i^{max} - f_i^{min}}$$

where $f_i^{max} = \mu + 2\sigma$, $f_i^{min} = \mu - 2\sigma$, and μ is mean feature value of all cases $(N = 964)$ and σ is the standard deviation. If $f_i^{Norm} > 1$, it is assigned to 1, while if $f_i^{Norm} < 0$, it is assigned to 0. In this way, we can avoid the possible impact by the outlier feature values in the dataset.

Next, the bilateral asymmetry features are computed using an absolute subtraction of two matched features extracted from the left and right breast of either CC or MLO views, independently (i.e., $f_i^{BS} = |f_i^{Norm-Left} - f_i^{Norm-Right}|$) to quantify bilateral breast tissue or image feature difference or asymmetry. Then, a variance thresholding method is applied to prescreen the compute bilateral asymmetrical features using an empirically selected threshold of 0.01 to remove irrelevant or redundant features. Thus, the final CC-HCR and MLO-HCR feature vectors are generated.

4.2.2.3 Deep Transfer learning feature extraction and reduction

To extract deep transfer learning (DTL) features directly from the image, we use a VGG16 network pretrained on the ImageNet database exactly as conducted in our previous study.[154] Because this network is pretrained on three channel color images from the ImageNet database, the network will take three channel images as an input. We create pseudo-color ROIs by stacking the original image, a bilaterally filtered image, and a histogram equalized image in the three channels and feed this image to the network. Details of creating pseudo-ROIs can be found in our previous work.[154] The previous studies have demonstrated that using pseudo-ROIs as inputs to the deep transfer learning model produce features that contain more relevant information for the prediction task than directly stacking the original ROI into the three channels.[116, 154] Since VGG16 takes a 224×224 image as an input, all ROIs of 64×64 pixels are resized using a bilinear interpolation.

The architecture of the VGG16 network is made up of five blocks, each of which contain either two or three convolutional layers followed by a max pooling layer after each convolution layer. Then, VGG16 network includes three fully connected layers. Since we use VGG16 network as an automated feature extractor, the top three fully connected layers are removed. As a result, 25,088 automated image features are extracted after the final max pooling layer and then normalized. The bilateral asymmetrical features are quantified in the same manner as the HCR features.

To significantly reduce the dimensionality of the extremely large sets of automated features, we take two steps. First, a variance thresholding method is applied to remove all features that have a variance of less than an empirically selected threshold of 0.02, which reduces the number of automated features from 25,088 to ~6,000. Second, we use a sequential forward feature selector (SFFS) algorithm implemented with a 4-fold cross-validation method wrapped inside a linear support vector machine (SVM) and using the area under the receiver operating characteristic curve (AUC) as an evaluation metric[116, 154] to obtain a final optimal CC-DTL and MLO-DTL feature vector. This feature selection method of using the SFFS algorithm has been applied and reported in our previous studies. [116, 154]

4.2.2.4 Classification and Model Evaluation

In this study, we investigate four different feature fusion strategies to determine which method produced the best results (**Figure 4-5**).

1. Method 1 is a feature level fusion followed by a two-stage classification system. In this method, for each projection view, the HCR and DTL feature vectors are first fused via concatenation. Next, the two fusion feature vectors are used to train two separate classifiers whose outputs are then fused and used to train a final classifier.

2. Method 2 is a two-stage classification system in which four separate classifiers are trained independently using either the HCR or DTL feature vector from either projection view. Then, the outputs of the four classifiers are fused and used to train a final classifier.

3. Method 3 is a three-stage classification system, which begins the same way as method 2. In the second stage, the outputs of the classifiers trained on the CC-HCR and CC-DTL feature vectors are concatenated and used to train one classifier and the outputs of the classifiers trained on the MLO-HCR and MLO-DTL feature vectors are concatenated and used to train another classifier. In the third stage the output of the two classifiers trained on either projection CC or MLO view are concatenated and used to train a final classifier.

4. Method 4 is identical to method 3 except in the second stage the two classifiers are trained using the concatenation of the outputs of the prior classifiers that were trained using either the CC-HCR and MLO-HCR feature vectors or the CC-DTL and MLO-DTL feature vectors.

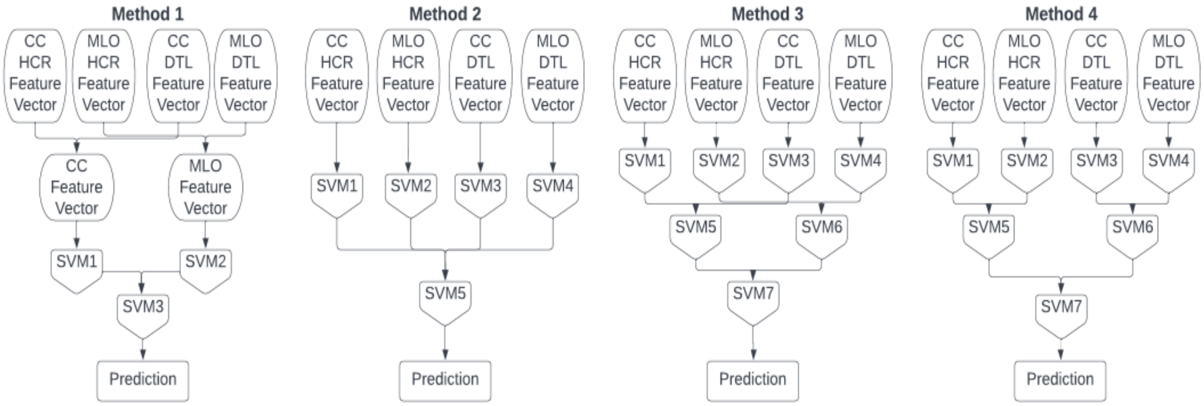

***Figure 4-5*:** Schematic Diagram of the four fusion methods investigated

As shown in **Figure 4-5,** we select a linear support vector machine (SVM) as the machine learning classifier to fuse image features and generate a classification score to predict the likelihood of a testing case depicting a malignant lesion because when comparing to many other types of machine learning classifiers, a SVM is easy to train with a simple structure and has a higher capability to be robust. Thus, SVMs are commonly used in CAD of breast lesion classification tasks as described in a recent systematic review article [50]. In this study, all SVMs are trained and tested using a stratified 10-fold cross validation method in which all cases were randomly divided into 10 subsets, where in each cross-validation cycle, nine subsets are used for training and one subset is used for the testing of SVM classifier. To address the imbalance issue of our dataset (36.6% benign cases versus 63.4% malignant cases, which will be reported in Results section below), we use the synthetic minority oversampling technique (SMOTE) to oversample the benign cases to ensure that each classifier is trained using a subset of the data that contains a balanced number of malignant and benign cases[155]. SMOTE algorithm is embedded into each cross-validation fold and applied to only the training datasets as reported in the previous study.[85]

Each SVM produces a prediction score between 0 and 1 for each testing case, where higher scores indicate a higher probability of being malignant. Prediction scores generated on the testing dataset over 10-fold CV are then used to generate receiver operating characteristic (ROC) curves using the publicly available ROC curve fitting program, ROCKIT (http://metz-roc.uchicago.edu/MetzROC), which generates a smooth ROC curve based on the maximum likelihood estimates of the SVM-generated prediction scores. The area under the ROC curve (AUC) along with the standard deviation is computed and used as an evaluation metric. The statistically significant difference of the different SVM classifiers (AUC values) are also compared using p-values computed by ROCKIT program. Additionally, an operation threshold of 0.5 is applied to the prediction scores to divide the testing cases into malignant and benign class. Predictions scores (<0.5) are classified as benign, while scores (≥0.5) are malignant. The overall classification accuracy, precision, sensitivity, and specificity of each SVM along with the standard deviation are then computed and recorded as additional evaluation indices.

4.3. Results

The initial dataset is comprised of 1,065 cases that contain four FFDM images, namely: LCC, RCC, LMLO, and RMLO images. Each case depicts one biopsied soft tissue mass-type breast lesion. Our ipsilateral matching scheme is unable to confirm that the lesion marked in the CC view is the same lesion marked in the MLO view in 66 cases. The bilateral registration scheme fails to register the other 35 cases. This resulted in a final dataset that contained 964 cases of which 353 cases depict benign lesions and 611 cases depict malignant as confirmed by tissue biopsy. Therefore, the final true case-based dataset used in this study to train and test the CAD scheme contains 3,856 FFDM images where 1,412 images associate with benign cases and 2,444 images associate with malignant cases.

After feature reduction, the HCR-CC and HCR-MLO feature vectors contain 26 and 22 features, respectively. The HCR features selected for the final feature sets are displayed in **Table 4-1**. The DTL-CC and DTL-MLO feature vectors contain 74 and 44 features, respectively. The HCR-CC feature vector and the DTL-CC feature vector are combined via concatenation to create the fusion feature vector that is used in Method 1.

The same process is repeated with the MLO feature vectors. To fuse HCR and DTL features, the features included in CC fusion and MLO fusion feature vectors are further analyzed and reduced using a SFFS method. After feature reduction, the CC fusion feature vector contains 43 features (including 14% HCR features and 86% DTL features) while the MLO fusion feature set contains 31 features (including 13% HCR features and 87% DTL features). The HCR features selected in the final fusion feature vectors are shown in the last two columns of **Table 4-1**.

Feature Type	**Feature Name**		**Feature Set**			
			HCR-CC	**HCR-MLO**	**Fusion (CC)**	**Fusion (MLO)**
Statistical	Mean		X	X		X
	Max		X	X		
	Standard Deviation		X	X		X
	Energy					
	Entropy		X	X		
	Skewness					
	Kurtosis		X	X	X	
GLCM	Contrast	Max				
		Mean	X			
	Dissimilarity	Max	X	X		
		Mean	X	X		
	Homogeneity	Max	X			
		Mean	X	X		
	ASM	Max				
		Mean				
	Energy	Max				
		Mean				
	Correlation	Max	X	X	X	
		Mean		X		
GLRLM	SRE	Max	X			
		Mean				
	LRE	Max				
		Mean				
	GLN	Max	X	X		
		Mean	X	X		
	RLN	Max				
		Mean				
	RP	Max				

		Mean				
	LGLRE	Max				
		Mean				
	HGLRE	Max	X	X		
		Mean	X	X		
	SRLGLE	Max	X			
		Mean	X		X	
	SRHGLE	Max	X	X		
		Mean	X	X		
	LRLGLE	Max				
		Mean				
	LRGHLE	Max	X	X		
		Mean	X	X	X	
Gabor Features	Mean		X	X		X
	Variance		X	X	X	X
	Energy		X	X		
	Entropy		X	X	X	

Table 4-1: Handcrafted radiomic features selected after feature reduction. An X indicates that the feature was selected to be used in the final feature vector for the corresponding column.

The results of the four different fusion methods are shown in **Figure 4-6** and **Table 4-2**, which include four ROC curves generated by four fusion methods (**Figure 4-6**) and the corresponding AUC values along with the standard deviation computed by ROCKIT program (**Table 4-2**). The results show that using Method 1, three SVMs yield significantly higher AUC values than the corresponding SVMs generated using methods 2, 3, and 4 (with all $p < 0.005$). The similar performance patterns (including classification accuracy, precision, sensitivity, and specificity) among the SVM classifiers generated in four methods are also observed after applying the operation threshold to assign or classify testing cases into malignant and benign classes. Thus, Method 1 is selected for further data analysis.

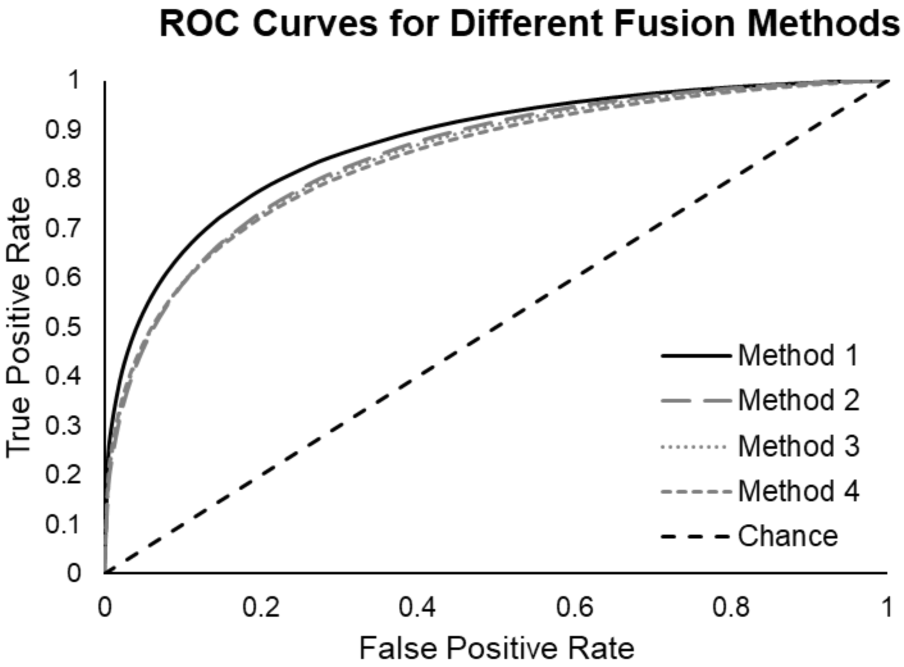

Figure 4-6: *Final ROC Curves of the four different fusion methods. ROC Curves are generated using a maximum likelihood estimation method in ROCKIT.*

Method	SVM	AUC	Accuracy	Precision	Sensitivity	Specificity
1	SVM1	0.817 ± 0.026	0.745 ± 0.033	0.745 ± 0.116	0.633 ± 0.057	0.841 ± 0.053
	SVM2	0.792 ± 0.026	0.721 ± 0.035	0.734 ± 0.048	0.600 ± 0.047	0.823 ± 0.027
	SVM3	0.876 ± 0.031	0.792 ± 0.044	0.773 ± 0.097	0.696 ± 0.059	0.863 ± 0.049
2	SVM1	0.664 ± 0.039	0.611 ± 0.030	0.694 ± 0.063	0.478 ± 0.027	0.763 ± 0.039
	SVM2	0.642 ± 0.051	0.584 ± 0.046	0.677 ± 0.047	0.456 ± 0.039	0.738 ± 0.041
	SVM3	0.781 ± 0.030	0.726 ± 0.023	0.714 ± 0.110	0.609 ± 0.027	0.823 ± 0.052
	SVM4	0.741 ± 0.029	0.694 ± 0.034	0.694 ± 0.069	0.572 ± 0.049	0.800 ± 0.029
	SVM5	0.851 ± 0.025	0.782 ± 0.030	0.748 ± 0.095	0.691 ± 0.053	0.850 ± 0.043
3	SVM1	0.664 ± 0.039	0.611 ± 0.030	0.694 ± 0.063	0.478 ± 0.027	0.763 ± 0.039
	SVM2	0.642 ± 0.051	0.584 ± 0.046	0.677 ± 0.047	0.456 ± 0.039	0.738 ± 0.041
	SVM3	0.781 ± 0.030	0.726 ± 0.023	0.714 ± 0.110	0.609 ± 0.027	0.823 ± 0.052
	SVM4	0.741 ± 0.029	0.694 ± 0.034	0.694 ± 0.069	0.572 ± 0.049	0.800 ± 0.029
	SVM5	0.800 ± 0.023	0.742 ± 0.042	0.714 ± 0.120	0.634 ± 0.049	0.825 ± 0.054
	SVM6	0.766 ± 0.032	0.709 ± 0.038	0.697 ± 0.072	0.595 ± 0.063	0.806 ± 0.032
	SVM7	0.852 ± 0.027	0.778 ± 0.035	0.742 ± 0.098	0.686 ± 0.057	0.846 ± 0.044
4	SVM1	0.664 ± 0.039	0.611 ± 0.030	0.694 ± 0.063	0.478 ± 0.027	0.763 ± 0.039
	SVM2	0.642 ± 0.051	0.584 ± 0.046	0.677 ± 0.047	0.456 ± 0.039	0.738 ± 0.041
	SVM3	0.781 ± 0.030	0.726 ± 0.023	0.714 ± 0.110	0.609 ± 0.027	0.823 ± 0.052
	SVM4	0.741 ± 0.029	0.694 ± 0.034	0.694 ± 0.069	0.572 ± 0.049	0.800 ± 0.029
	SVM5	0.642 ± 0.051	0.581 ± 0.046	0.657 ± 0.033	0.452 ± 0.039	0.728 ± 0.036
	SVM6	0.829 ± 0.028	0.762 ± 0.029	0.734 ± 0.090	0.659 ± 0.037	0.838 ± 0.042
	SVM7	0.841 ± 0.028	0.772 ± 0.033	0.742 ± 0.057	0.676 ± 0.056	0.842 ± 0.028

Table 4-2: *Results of the four fusion methods. Mean values and standard deviation over 10-fold CV.*

We further analyze the data listed in **Table 4-2**. First, to further analyze the differences between feature level fusion and output level fusion, we compare the performance of classifiers of method 1 to stage two classifiers of method 3. In method 1, SVM 1 and SVM 2 are trained using a feature vector that fuses HCR and DTL feature vectors computed from the CC view and MLO view, respectively. In method 3, SVM 5 and SVM 6 are trained using the fusion of the outputs from classifiers independently trained on the HCR and DTL feature vectors computed from the CC and MLO views, respectively. We compare the performance between SVM 1 of Method 1 and SVM 5 of Method 3, as well as between

SVM 2 of Method 1 and SVM 6 of Method 3 to determine if direct fusion of features computed from multi-view images continues to outperform fusion of classifier output scores generated by multi-classifiers trained only using image features computed from a single projection view. The data analysis results show that SVM 1 and SVM 2 of method 1 yield significantly higher AUC values (AUCs = 0.817±0.026 and 0.792±0.026) than SVM 5 and SVM 6 of Method 3 (AUC = 0.800±0.023 and 0.766±0.032) with $p = 0.0327$ for using two bilateral CC view images and $p < 0.001$ for using two bilateral MLO view images, respectively, which indicate that fusion of image features is better than fusion of output of two classifiers separately trained using different single-view image features.

Second, to determine whether fusion of HCR and DTL feature vectors yield better results, we compare the performance of the SVMs trained on the fusion feature sets used in Method 1 to the SVMs trained on the HCR and DTL feature sets independently in stage one of all three output level fusion methods. For both projection views, the SVMs trained using the HCR and DTL fusion feature vectors also yield significantly higher classification performance (AUCs = 0.817±0.026 and 0.792±0.026) than the SVMs trained using either only the HCR or DTL feature vector (AUCs = 0.664±0.039 and 0.781±0.030) with $p < 0.001$ and $p = 0.0431$ for the CC view, and (AUCs = 0.642±0.051 and 0.741±0.029) with $p < 0.001$ and $p = 0.0091$ for the MLO view, respectively.

Third, besides that the SVMs of Method 1 in general perform significantly better than the SVMs of the other three methods, we also compare the performance between the SVM trained using four images that combine two pairs of bilateral images (CC and MLO view) and other two SVMs trained using two images that combine one pair of bilateral images (either CC or MLO view), which are SVM1 vs SVM2 and SVM1 vs SVM3 as shown in Method 1 of **Figure 4-5**). The results show that SVM1 yields an AUC = 0.876±0.031, which is significantly higher than AUC = 0.817±0.026 yielded by SVM2 and AUC = 0.7920±0.026 yielded by SVM3 (both p < 0.001) **(Table 4-2).** Corresponding ROC Curves are displayed in **Figure 4-7.** No statistically significant difference is observed in the ROC curves or AUC values between SVM2 and SVM3 (p = 0.3546). Additionally, **Figure 4-8** displays the sum of three confusion matrices of SVM1, SVM2 and SVM3 computed based on the classification accuracy of malignant and benign cases, which are then used to compute the overall classification accuracy, precision, sensitivity, and specificity as reported in **Table 4-2**.

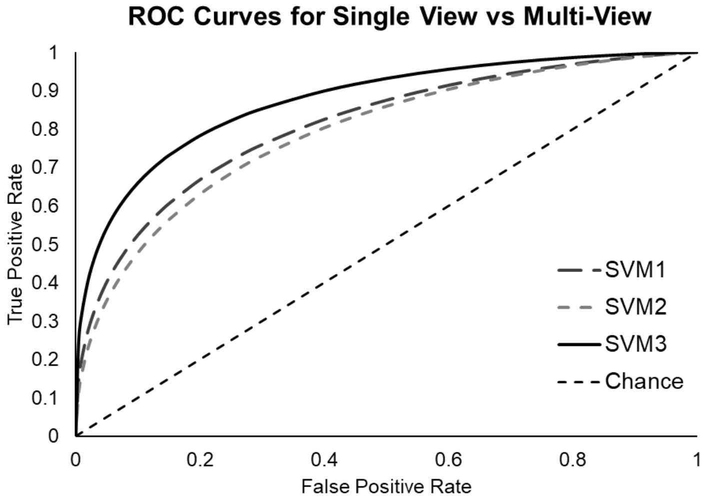

Figure 4-7: *smooth ROC curves of the single-view classifiers and the multi-view classifier based on the maximum likelihood estimates of the prediction scores generated over 10-fold CV.*

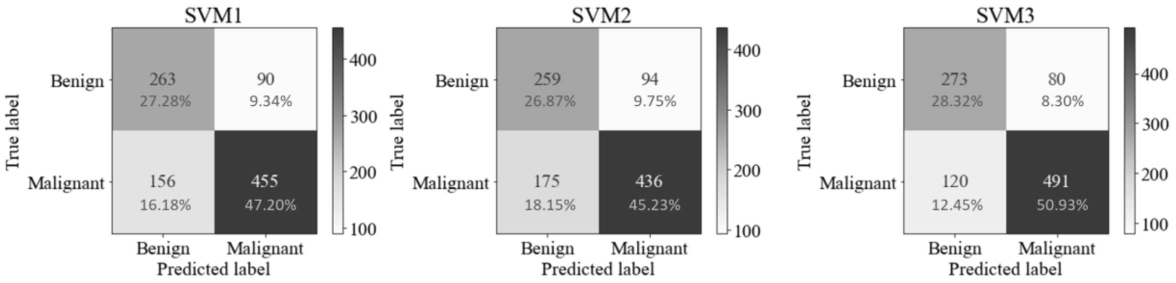

Figure 4-8: *Sum of each confusion matrix over 10-fold CV for the single view and multi-view classifiers.*

4.4. Discussion

This paper reports on a new study that combines three common analysis tools used in developing CAD of multi-view mammograms into a single framework for assisting in the diagnosis of suspicious breast lesions as malignant or benign. Unlike previous multi-view CAD schemes of mammograms that combine the complementary image features computed from either ipsilateral or bilateral mammography views, or the CAD schemes the use both HCR features and DTL features computed from single images, this study has several unique characteristics or aspects as comparing to many previous studies in this research field.

First, this is a complete case-base CAD scheme that extracts two sets of matched ROIs from four mammograms in one screening examination (including two lesion regions depicting on two ipsilateral views and two negative regions on images of the contralateral breast). Two types of image features (HCR and DTL) computed from these four matched ROIs are passed through the framework simultaneously, so that the final machine learning classifier fuses the clinically relevant and complementary information extracted from each ROI of different view when making a final predictive decision. However, accurate identification of four matched ROIs on both ipsilateral and bilateral mammograms by a CAD scheme is very difficult due to the difference of breast compression in acquiring four view images. Unlike previous studies (i.e., Khan et al.[84]) that manually determine four matched ROIs from four mammograms, we develop and add two algorithms of an ipsilateral view matching and a bilateral image registration prior to ROI extraction. As a result, applying an ipsilateral matching algorithm ensures that one lesion visualized in one projection view is the same lesion visualized in another projection view. This is an extremely important step as some cases may have a suspicious lesion marked in the CC view and a different lesion marked in the MLO view, meaning there are two distinct lesions within the breast, and each is only visualized in one projection view. Additionally, our CAD framework also applies a bilateral registration algorithm to ensure that the ROIs extracted from the contralateral breast are from the same spatial location as the lesion on two ipsilateral view images. By implementing these two algorithms, we developed a unique four-view image or case-based CAD framework.

Second, we chose to quantify the bilateral asymmetrical features computed from two ROIs in each projection (CC and MLO) view as opposed to using the image features computed from two bilateral ROIs independently to build and train machine learning classifier. Our approach does not only reduce the number of image features in the initial feature pool, which improve efficacy of feature selection or feature dimensionality reduction, it can also better mimic the experience of how radiologists diagnose breast lesions in reading mammograms. Since when visually inspecting a mammogram exam for abnormalities, a radiologist often relies on the bilateral asymmetry as a qualitative imaging marker, quantifying bilateral asymmetry of two pairs of the matched ROIs in CC and MLO view can also generate effective quantitative imaging markers used in CAD schemes. Previous studies have demonstrated the advantages of applying CAD schemes that focus on analysis of bilateral image feature asymmetry computed from two mammograms of left and right breast to predict the short-term risk of developing breast cancer[89-91, 156] and the likelihood of having breast cancer depicting on mammograms.[85], [86] However, these previous studies bypass the image registration step and the extraction of ROIs. Thus, the prediction models or classifiers are developed based on the analysis of bilateral image feature asymmetry computed from whole breast. Our study computes bilateral image feature asymmetry two matched ROIs, which can eliminate or significantly reduce the impact of the most heterogeneously normal breast tissue areas, and thus help improve CAD performance of lesion classification. For example, one previous CAD scheme using bilateral image feature asymmetry of whole mammograms reported a macro-AUC of 0.733 in detecting breast cancer,[86] while our CAD scheme yields AUC = 0.876±0.031. Although two studies use different image datasets and their performance cannot be directly compared, we believe that classification performance of our new CAD scheme is encouraging, which is attributed to the quantification of bilateral image feature asymmetry of the targeted ROIs matched in pairs of bilateral mammograms.

Third, unlike many previous CAD schemes that use either traditional HCR features or automated DTL features separately, this study demonstrates the feasibility and advantages of fusing HCR and DTL features computed from two pairs of bilateral ROIs extracted from four mammograms. In using the pretrained VGG16 network as a feature extractor, we are able to mix HCR and DTL features into one initial feature pool. Thus, the optimal fusion feature vectors include both HCR and DTL features, which provide lower correlation or complementary information. Additionally, we also observe that in fusion feature vectors, majority of features are DTL feature (i.e., CC fusion feature vector contains 6 HCR features (14%) and 37 DTL features (86%)), which shows that DTL features make higher contribution in this CAD scheme. However, adding the minority of HCR features still improves classification performance of the final fusion-based CAD scheme. In addition, although several other studies have also been conducted to fuse HCR and DTL features to develop CAD schemes of breast lesion classification, these schemes are limited to be single-view or faux case-based schemes as ROIs are extracted from all four views and classified independently or from only two-views omitting information contained in the contralateral [41, 79]. Our study is the first study that fuses the bilateral asymmetry of HCR and DTL features computed from two pairs of the matched ROIs on CC and MLO views.

Fourth, although many fusion methods have been previously investigated aiming to help improve CAD performance, few studies have investigated and compared different fusion methods to identify the optimal method for the multi-level fusion problem. In this study, we test three fusion methods or tasks in developing this CAD framework namely, bilateral image fusion, ipsilateral image fusion, and finally fusion of multiple feature types. The first level of fusion is handled through the quantification of the bilateral asymmetry as this is when we fuse information extracted from two bilateral mammograms. Our justification for this type of fusion is based on the location of bilateral asymmetry in mammograms as an indicator of abnormalities. To determine the optimal way to fuse ipsilateral information and multiple feature types, we conducted several experiments with four different fusion methods. Results show that feature level fusion of the different feature types prior to training classifiers on each projection view is superior to output level fusion after training classifiers on each feature set independently.

Due to the above unique characteristics or innovation of this study, we also make several interesting observations to further support or validate several important conclusions of previous studies. First, in our previous work, we developed a single view CAD scheme that fused HCR and DTL features extracted from only the CC view of a lesion and concluded that the CAD scheme trained by fusion of HCR and DTL features could yield significantly higher performance than the CAD schemes developed using only either HCR or DTL features.[157] This work is an extension of our previous study which solidifies this conclusion using both the CC and MLO projection views. Second, we observe that late fusion of information extracted from different projection views performs better than when this information is fused earlier. This can be seen by the results of method 1 and method 3 as both methods keep the two projection views separate until the final classification step and yield the best classification performance in terms of AUC. We believe that this is because fibroglandular tissue patterns often appear very different on CC and MLO view projection images, which makes the information contained in the feature vectors extracted from the two view images very different. Hence, the superior result is obtained when the information extracted from multiple projection views is used to train classifiers separately. Third, we also observe that multi-view CAD systems tend to outperform single view CAD systems as demonstrated in many previous studies.[88, 157] This conclusion is further validated and expanded in this study using a combination of HCR and DTL features from all four view mammograms in a complete case-based manner. In this study, the four-view fusion CAD system yields a classification performance of AUC = 0.876±0.031 with an accuracy of 0.792±0.044, while the performance of CAD schemes based on fusion of two bilateral images of either CC or MLO view only yield AUC of 0.817±0.026 and 0.792±0.026, and an accuracy of 0.745±0.033 and 0.721±0.035, respectively.

Although this is a unique case-based multi-view CAD framework that yields an encouraging performance of breast lesion classification, we recognize the limitations of this study. First, we use a relatively simple ROI extraction technique to avoid introducing any potential bias or variability from an automated or semi-automated tumor segmentation scheme. This method may not have been optimal, therefore, we should investigate other lesion segmentation techniques prior to feature extraction.

Second, although many deep learning models have been used in CAD field as feature extractors, we used a pretrained VGG16 network as a feature extractor to decrease the computational complexity of this framework since using transfer learning for feature extraction does not require additional training of the network. We should test and compare different networks and methods for extracting the DTL features from these deep networks (i.e., using another popular ResNet50 model in CAD schemes[157]). Additionally, we plan to investigate the effects of transfer learning using a DL network pretrained on radiological images from the RadImageNet database as opposed to the natural images in the ImageNet database, as RadImageNet pretrained models have outperformed ImageNet pretrained models in some medical classification tasks. [36]

Third, we conduct the feature reduction and selection process to identify the optimal feature vectors using the whole dataset. To minimize the possible bias, we also apply a 4-fold cross validation method in feature selection as reported in previous CAD studies.

[37] Then, the features are used to build SVM classifiers using a 10-fold cross-validation method. Although this approach has advantages of identifying the final optimal feature vectors, it may introduce the risk of increasing bias to the classifiers because the testing cases are only blind to classifier training process and may be involved in feature selection process. To eliminate the possible bias, the feature selection process should be embedded into the cross-validation of classifier training and testing process, however, this process has disadvantages of higher computation costs and the inability to identify the final optimal feature vectors that can be applied "as is" to the new independent datasets in future validation studies. For this study, we believe that the impact of the potential bias can be ignored because our objective is to compare the relative performance changes among the several SVMs that are built using the same feature selection and classifier training and testing method.

Fourth, this framework is developed and tested using a singular dataset, therefore, it may not be generalizable to other mammography images that were taken at different centers on different machines. To further test and improve the generalizability and robustness of this new CAD framework, we will continue to expand our study dataset by collecting new images from our university medical center and utilizing publicly available databases in our future studies.

Fifth, we recognize that in current clinical practice, more and more 2D mammograms are synthetic images generated by digital breast tomosynthesis (DBT) images, which may have slightly different image quality or characteristics as comparing to original FFDM images. Thus, CAD scheme developed using FFDM images may need to be retrained to fit the DBT-generated synthetic images. However, the approved concept of this study is also valid to the DBT-generated synthetic images.

Last, this study only includes soft-tissue mass type lesions seen on both projection (CC and MLO) views, while this excludes a small fraction of subtle or difficult lesions. Future work should include cases where a mass is only seen in one view by developing and adding a new CAD module to handle and process these difficult cases.

4.5. Conclusions

In summary, we develop and test a novel case-based CAD framework of breast lesion classification in this study, which (1) extracts two sets of matched ROIs from the CC and MLO view of mammograms, (2) computes a set of bilateral asymmetric HCR and DTL image features (3) assembles two optimal fusion feature vectors mixed with both HCR and DTL features, and (4) builds final machine learning classifier (SVM) trained using the fusion feature vectors. By applying this new CAD framework to a diverse image dataset involving 964 cases of 3,856 FFDM images, we conduct a series of experiments to compare advantages and lesion classification performance using different image feature or classification score fusion methods. The study results demonstrate that (1) fusing HCR and DTL features for each pair of projection view before training a classifier is a better choice than fusing the outputs of classifiers trained on each type of features independently and (2) CAD classification performance is enhanced through the addition and fusion of image features computed from two ipsilateral (CC and MLO) views of the lesion. Overall, the study results fully support our hypothesis that (1) HCR and DTL features contain complementary information in lesion classification, (2) multi-view CAD outperforms single-view CAD for mammography lesion classification. The study results also highlight the significance of optimally fusing HCR and DTL image features computed from all four matched mammograms to enhance performance of the final CAD classifiers. However, this is a proof-of-concept type study, more work needs to further optimize and validate this new case-based CAD framework in future studies.

Chapter 5. Pseudo color image generation for improving the performance of deep transfer learning-based computer aided diagnosis schemes in breast mass classification

5.1. Introduction

Breast cancer is one of the leading causes of death in women worldwide. While the mortality rate of breast cancer has dropped 42% since 1989, the incidence rate of breast cancer continues to increase by 0.5% each year. [1] Mammography, a population-based x-ray screening tool, plays a large role in these statistics as it helps with early detection which is key for keeping the mortality rate low. In a standard mammography screening exam, two images are taken of each breast, a craniocaudal (CC) view which is taken from the top, and a mediolateral oblique (MLO) view which is taken from the side. Radiologists will analyze all four images (two projection views from both breasts) to determine if there are any suspicious regions that must be biopsied. Even though mammography has played a significant role in decreasing the breast cancer mortality rate, there is a very high false positive rate associated with the exam as less than 30% of suspicious regions referred for biopsies are malignancies[100]. This is because mammography images are very difficult to interpret due to high heterogeneity between lesions and difficulty associated with visualizing dense breast tissues.

Many computer-aided diagnosis (CADx) systems have been developed which aim to help a radiologist classify suspicious lesions and thus decrease the false positive rate. These systems can act as a second reader which can decrease the workload on radiologists as well as decrease the amount of time spent analyzing each exam. However, the utility and effectiveness of the systems used in clinical practice is often questioned as there are conflicting results as to whether these systems really aid in decreasing the false positive or benign biopsy rates[158, 159]. Much more work must be done to make these systems more robust in addition to determining the best way to fuse these systems into the clinical workflow. Despite that caveat, the utility of artificial intelligence into experimental CADx schemes for mammography has allowed for tremendous progress in the medical imaging field.

Deep learning based CADx schemes use convolutional neural networks (CNNs) to automatically classify a suspicious lesion from the input image. Since these networks learn the image features directly from the suspicious lesion, they can identify patterns relevant to the target domain that cannot be seen with the human eye. While these networks tend to outperform machine learning based CADx schemes, they are more difficult to train as they require a large and diverse dataset which is not often available in the medical imaging domain[34, 160]. Transfer learning has emerged as a solution to this problem. Transfer learning is a method in which a network that has been trained on a large dataset is modified and used for a different task. This works well because of the deep structure of the CNNs; the initial layers of the network are able to learn generic, high-level features, where the deeper layers can learn features more specific to the target domain. Transfer learning has been used extensively in many breast cancer classification tasks by either fine-tuning the network or as a feature extractor, both techniques which have shown promising results[65].

Transfer learning studies often use a deep CNN pretrained on the ImageNet dataset. The ImageNet dataset is comprised of three-channel RGB color images that tend to have a single focal point[61]. Mammogram images are single channel greyscale images which do not always have a distinct focal point, especially when looking at very dense breast tissue. Despite these differences, transfer learning using a state-of-the-art network such as VGG16, ResNet50, or InceptionV3, pretrained on ImageNet has still had tremendous success in breast mass classification tasks[50]. However, using this method requires our input mammography images to also be transformed into three channel images to match the shape of the ImageNet images. This is termed pseudo color image generation as the single channel images are transformed into pseudo RGB images to be compatible with the pretrained network. Many studies do not discuss the method in which this is conducted. Most commonly, we see the mammogram images stacked in three channels before being fed to the deep network. There are a limited number of studies that begin to investigate the potential benefits of different methods of pseudo color ROI generation.

Razali et al. created pseudo color images by mapping the single-channel greyscale ROI to an RGB color map. Using a pretrained ResNet50 network as a feature extractor and an SVM for classifying suspicious lesions as malignant or benign, they demonstrated that color manipulation of the original greyscale images provides increased information and can yield better performance than using greyscale images alone (ACC of 91.54 vs 88.56, respectively) [92]. Teare et al. utilized pre-processing techniques in the pseudo color image generation step to increase the feature representation of the input images. Pseudo color images were developed by varying the CLAHE window and clipping parameters in the three channels. An inceptionV3 network was pretrained on ImageNet and used as a feature extractor and input to a random forest for classification which resulted in an AUC of 0.922, specificity of 0.80, and sensitivity of 0.91 [93]. Jones et al.

compared the performance of two different pseudo color image sets for classifying breast lesions in the craniocaudal view as malignant or benign. The first was created by stacking single-channel greyscale images in three channels, and the second by stacking the original greyscale image, a bilaterally filtered image, and a histogram equalized version. A VGG16 network pretrained on the ImageNet dataset was used as a feature extractor. DTL-based features were merged with handcrafted features extracted from the original greyscale images and fed to a SVM for predicting the likelihood of malignancy. Better performance was seen using the pseudo color ROIs generated using pre-processing techniques compared to the ROIs created via stacking (AUC 0.756 vs 0.734 and Accuracy 0.704 vs 0.676, respectively) [154].

Li et al. compared the performance of a CNN in detecting masses using two different pseudo color inputs. The first dataset used pseudo color ROIs created by stacking the original ROI, a gradient image, and the local ternary pattern image. The second dataset uses the single channel greyscale ROI images. The free receiver operating characteristics (FROC) curves did not show any statistically significant differences in the ability to distinguish masses from normal tissue using pseudo color ROIs or greyscale ROIs. Authors assert that this may be because the pseudo color ROI was generated using gradient and texture images that can be learned by the CNN itself, therefore creating the pseudo color image in this manner may not be actually increasing the amount of information at all [94]. Min et al created pseudo color ROIs by placing the original image in the red channel, followed by two images generated by a multi-scale morphological sifter (MMS) in the green and blue channels. The MMS was developed in a manner that aimed to extract spicules as malignant lesions tend to have spiculated margins. The pseudo color ROI is then used in a pretrained Mask R-CNN network with a ResNet101 backbone for mass detection and segmentation. Using pseudo color images outperformed the single channel ROIs with an average true positive rate of 0.90 at 0.9 false positives per image and an average dice similarity index of 0.88 [95].

These studies demonstrate that pseudo color image inputs may provide increased and complementary information to a CNN which yields better classification and detection performance. Therefore, the goal of this study is to further investigate the effects of using different pseudo color images as input to a pretrained deep CNN for classifying suspicious breast lesions and malignant or benign. It is known that malignant and benign masses can often be distinguished based on their contours as benign masses tend to have round or oval contours while malignant masses tend to be irregular in shape with highly spiculated margins[32, 40, 41]. We hypothesize that the addition of a fully segmented mass that captures these morphological distinctions to a pseudo color image may aid in classifying suspicious breast lesions as malignant or benign by providing increased and complementary information. In this study, we compare the performance of seven pseudo color image sets as inputs to a multi-view CADx scheme for classifying suspicious breast lesions as malignant or benign. Section 2 of this paper details the methods used in this study, including the pseudo color image generation steps and the mass segmentation and classification frameworks. The results are presented in section 3 followed by a discussion of the results in section 4.

5.2. Methods

5.2.1. Dataset

The dataset used in this study consists of full-field digital mammograms (FFDM) that were acquired under institutional review board (IRB) approved protocol using a Hologic Selenia digital mammography machine (Hologic Inc., Bedford, MA USA) from 2008 to 2014. Details pertaining to the image specifics can be found in several of our previous publications[36, 154, 161]. In this study, we retrospectively assembled a dataset by selecting cases that contained all four images taken during the screening mammography exam (left and right CC and MLO images). We only select cases where a mass can be seen in both the CC and MLO view of the same breast. An experienced radiologist has marked the center of each lesion and all suspicious lesions have been biopsy proven as malignant or benign. The center of each lesion was used as a reference to manually annotate the suspicious mass boundaries. The annotations were converted into binary segmentation masks and treated as the ground truth.

5.2.2. Extraction of matched ROIs

As some cases contain multiple masses in the same breast, we conducted an ipsilateral matching scheme to ensure that masses were properly matched up prior to classification. After ipsilateral matching, each mass is represented by four images, the CC and MLO view of the breast that contains the suspicious mass and the CC and MLO view of the contralateral breast. In this study, we will quantify the bilateral asymmetry between breasts as conducted in our previous work. To do so, we first must register the contralateral breast images to the image that contains the suspicious mass. The bilateral registration scheme is conducted using a multi-resolution B-spline transformation that optimizes the mattes mutual information metric. The FFDM image containing the suspicious mass is selected as the fixed image and the contralateral is selected as the moving image so as to not distort the location of the mass that has been marked by a radiologist. In depth details pertaining to the ipsilateral matching and bilateral registration schemes can be found in our previous study[162]. After bilateral registration, a 64x64 pixel region of interest (ROIs) is extracted surrounding the center of the lesion that has been marked by a radiologist. Since only the CC and MLO images of the breast containing the lesion have been center marked, we extract the same ROI from the contralateral breasts as the images have been bilaterally registered therefore, they should represent the same area in the breast. A visual representation of this can be seen in **Figure 5-1** step 1. In this figure, the suspicious lesion can be seen in the red bounding boxes in the left breast.

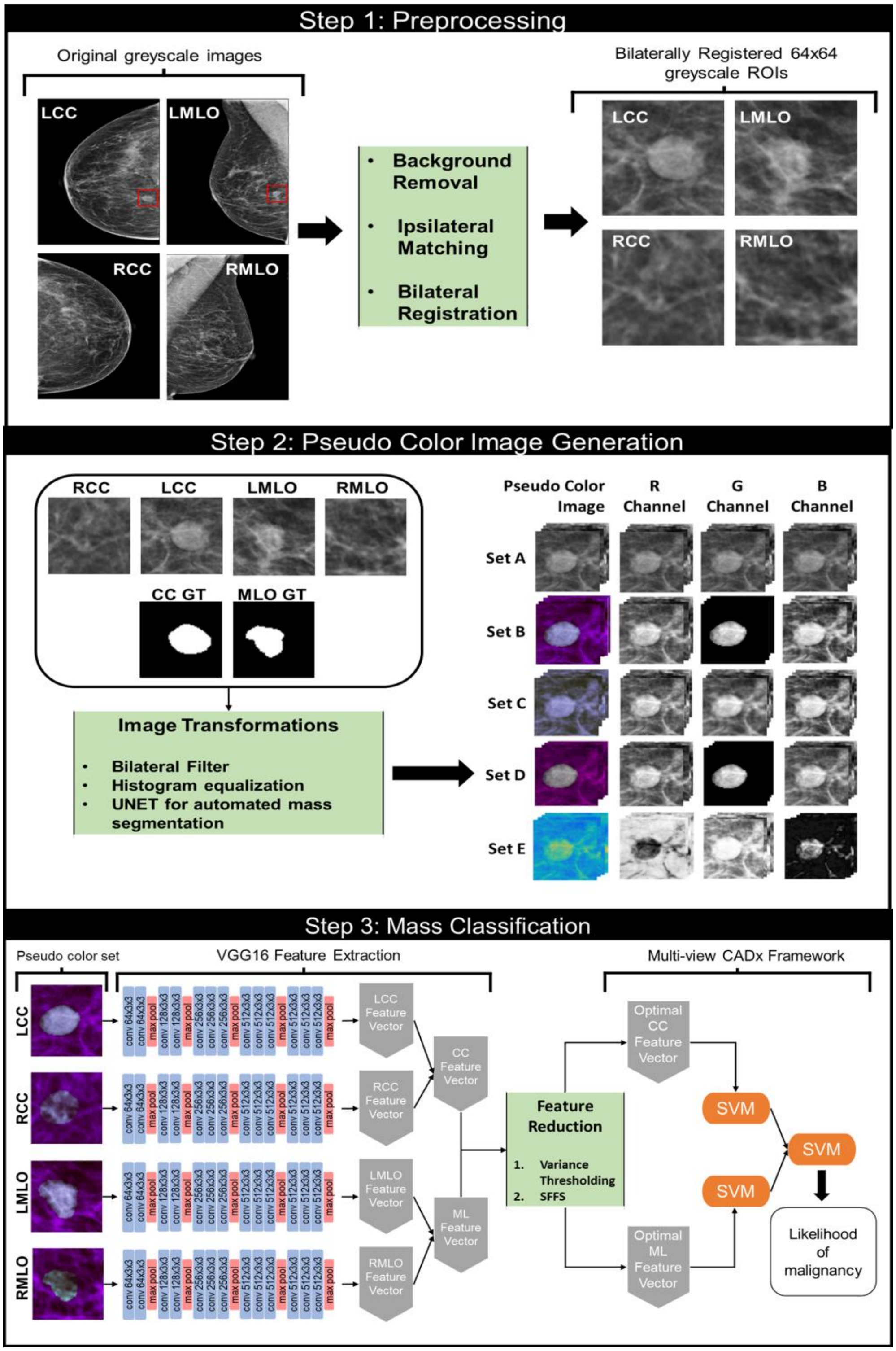

Figure 5-1: *A schematic diagram of this study.*

5.2.3. Pseudo color image generation

The aim of this study is to investigate the effects of different pseudo color ROIs on suspicious breast lesion classification. We create seven variations of pseudo color ROIs by stacking different versions of the original ROI in three channels, thus creating three channel images that mimic the three channel color images in the ImageNet dataset. **Table 5-1** contains a breakdown of each pseudo color image set used in this study. In all but one set, we use a combination of the original single-channel image (I_o), a histogram equalized version (I_{HE}), a bilaterally filtered version (I_{BF}), and the segmented mass image (I_{seg}). The final set is created by mapping the single-channel greyscale ROI values to all available values in the parula colormap. The histogram equalization pre-processing technique is selected to increase the contrast as mammography is an x-ray-based technique that is inherently low contrast. The bilateral filtering pre-processing technique is selected to de-noise the images as this technique is able to preserve edge and textural information while reducing noise [118].

Set	**R**	**G**	**B**
A	I_o	I_o	I_o
B_gt	I_o	I_{seg_GT}	I_{HE}
B_unet	I_o	I_{seg_unet}	I_{HE}
C	I_o	I_{BF}	I_{HE}
D_gt	I_o	I_{seg_GT}	I_o
D_unet	I_o	I_{seg_unet}	I_o
E	*Parula color mapped*		

Table 5-1: *Descriptions of each Pseudo color image sets. I_o = original image, I_{HE} = histogram equalized variant, I_{BF}= bilaterally filtered variant, I_{seg_GT}= segmented mass using the ground truth, I_{seg_unet}= segmented mass using the UNET predicted mask. Set E is created by applying the parula colormap to the original image.*

To best investigate the effect of including a segmented mass image in the pseudo color images, we use both the ground truth segmentation mask and a UNET generated automated segmentation mask to create the pseudo color images. The motivation for using an automated segmentation step is to bypass the need for a manual segmentation step in future studies as this is extremely time consuming and very user dependent[96]. Investigating the performance of both the ground truth segmentation as well as an automated segmentation result in pseudo color image will allow us to determine to what degree the automated segmentation result affects mass classification performance. Therefore, Set B and Set D are divided into two different sets, Set B ground truth(B_gt), which contains the ground truth segmentation, and Set B Unet (B_unet), which contains the UNET predicted segmentation mask. The same follows for Set D. This makes for a total of 7 pseudo color sets being tested. The fully segmented mass images for sets B_gt and D_gt are obtained by converting the ground truth annotation to a binary image and then applying this mask to the original greyscale image.

5.2.3.1 UNET for automated mass segmentation

Prior to the creation of the pseudo color ROIs, a Unet was trained to perform automatic mass segmentation. The Unet is selected as this segmentation method has been shown to perform well in breast mass classification tasks[34, 163-165] and tends to perform better than other segmentation methods like SegNet or Fully Convolutional Networks (FCNs) with a limited number of training examples[31]. The Unet is made up of an encoding pathway and a decoding pathway. The encoding pathway learns abstract representations of the input image through convolutional operations and downsampling via max pooling layers. The decoder pathway uses the abstract information learned from the encoder blocks to reconstruct the segmentation mask of the input via upsampling and convolutional blocks. The network is able to do this through skip connections which act as a bridge between the corresponding encoder and decoder blocks, allowing spatial information to be preserved[166]. The encoding pathway and decoding pathway form two symmetrical halves of the network which give the network its “U” structure.

The encoding half of the network consists of four blocks, each block contains two convolutional layers with a 3x3 kernel and ReLU activation functions, followed by a 2x2 max pooling layer with a stride of 2 for the downsampling. Each step down in the encoding pathway doubles the number of feature channels. The decoding half also consists of four blocks, each block upsamples the input feature map using a 2x2 up-convolution, then concatenates the feature map with the feature map from the corresponding encoding layer, followed by two convolutional layers with a 3x3 kernel and ReLU activation functions. Each step up in the decoding pathway halves the number of feature channels. All convolutional operations are padded with a stride of 1 to ensure the input and output image are the same size. A dropout layer with probability 0.2 is added to each block in the encoding layer and decoding layer after the first convolutional layer. The output layer of the network uses a 1x1 convolution with a sigmoid activation function to get the final probability map of each pixel belonging to either the foreground or background. To train the network, Adam optimizer was used with an initial learning rate of 1e-3 on mini batches of size 16 for 100 epochs. A custom loss function that combines binary cross entropy loss and dice loss was used to emphasize capturing fine details such as spiculations (**equations 5-1 – 5-3**).

$$L_{combined}(y_{true}, y_{pred}) = \alpha L_{Dice}(y_{true}, y_{pred}) + (1-\alpha) L_{BCE}(y_{true}, y_{pred}) \quad (5\text{-}1)$$

$$L_{Dice}(y_{true}, y_{pred}) = 1 - \frac{2 * |y_{true} \cap y_{pred}|}{|y_{true}| + |y_{pred}|} \quad (5\text{-}2)$$

$$L_{BCE}(y_{true}, y_{pred}) = -(y_{true} * \log(y_{pred}) + (1 - y_{true}) * \log(1 - y_{pred})) \quad (5\text{-}3)$$

Where y_{true} is the binary ground truth image, y_{pred} is the predicted binary segmentation mask, and α is a weight given to each component in the loss term. Different values for α were investigated. The final value is set at 0.7. The manual segmentations are treated as the ground truth. The learning rate decayed exponentially at a rate of e^{-1} after the first 10 epochs. The images containing masses are split into five bins, each bin is used as the testing set once for a total of five training cycles. In each training cycle, 80% of the data is used for training (60% for training, 20% for validation), while 20% is reserved for testing. This is done in a manner so that each image appears in the testing set once, therefore each image will have a corresponding segmentation mask produced by the UNET which can be used to create the pseudo color image. To create the segmented mass image, the binary mask created by the UNET will be applied to the original image to produce an image that contains only the segmented mass.

While there are four images associated with each mass, only two of the four images contain the suspicious mass. To create pseudo sets B_unet and D_unet, the binary segmentation mask will be applied to the contralateral image of the same projection view. For example, if the mass can be seen in the LCC and LMLO images, then the segmentation mask for the LCC image will be applied to the RCC image and the LMLO segmentation mask will be applied to the RMLO image. Since the contralateral breast images have been bilaterally registered to the images that contain the mass, applying the binary segmentation masks in this manner will segment out the area in the contralateral breast that correspond to the same location that the suspicious mass is located in. To assess the performance of the mass segmentation, we used the dice similarity coefficient (DSC) and the Jaccard Index

5.2.4. Mass Classification

Recent attention to CADx in breast cancer has demonstrated that multi-view CADx schemes tend to outperform single-view CADx schemes[84, 88]. Our previous study investigated the optimal method for fusing feature vectors from multiple views taken during a mammography exam[162]. We follow the same pipeline in this study. This involves extracting deep transfer learning-based features from each of the four mammography views, quantifying the bilateral asymmetry, and training a two-stage classification system which predicts the likelihood of a mass being malignant.

5.2.4.1 Transfer learning and feature extraction

In this work, we take advantage of the publicly available VGG16 network that has been pretrained on the ImageNet database. Before using the pretrained network, our input images are modified to match the ImageNet input image shape of 224x224x3. The 64x64 pixel original images are resized to 224x224 via bilinear interpolation and converted into three channel pseudo color images as mentioned in section 2.2. The VGG16 network has a relatively simple architecture that consists of five convolutional blocks followed by some dense layers. Blocks 1 and 2 contain two convolutional layers followed by a max pooling layer, while blocks 3, 4, and 5 contain three convolutional layers followed by a max pooling layer [124]. In this study, we use the pretrained VGG16 network as a feature extractor by freezing all weights and removing the top dense layers. The architecture adopted for this study can be seen in **Table 5- 2**. Features are then extracted from all four images after the final max pooling layer and flattened into a 25,088-dimensional vector.

Block	Layer	Size	Filter Size
1	Convolution-1	224×224×64	3×3
	Convolution-2	224×224×64	3×3
	Max pooling	112×112×64	-
2	Convolution-1	112×112×128	3×3
	Convolution-2	112×112×128	3×3
	Max pooling	56×56×128	-
3	Convolution-1	56×56×256	3×3
	Convolution-2	56×56×256	3×3
	Convolution-3	56×56×256	3×3
	Max pooling	28×28×256	-
4	Convolution-1	28×28×512	3×3
	Convolution-2	28×28×512	3×3
	Convolution-3	28×28×512	3×3
	Max pooling	14×14×512	-
5	Convolution-1	14×14×512	3×3
	Convolution-2	14×14×512	3×3
	Convolution-3	14×14×512	3×3
	Max pooling	7×7×512	-

Table 5-2: *VGG16 architecture used for feature extraction.*

5.2.4.2 Multi-view CADx Framework

A multi-view CADx framework is used to classify each case as malignant or benign. After pseudo color image generation, there are seven image sets. Each set contains four subsets representing the left and right CC and MLO image of the mass. Each subset contains m images of size 224x224x3, where m is the number of cases in this study. For each set, features are extracted from each subset independently, resulting in four m x 25,088 feature vectors. Each feature vector is then normalized from 0-1, and the bilateral asymmetry is quantified by taking the absolute value of the different between the feature vectors representing the bilateral views (LCC-RCC, LMLO-RMLO). This results in two feature vectors, a CC and MLO feature vector, each of size m x 25,088. To reduce the high dimensionality of these vectors, we first apply variance thresholding at a threshold of 0.2 to each set. We then apply a sequential forward feature selection (SFFS) algorithm to obtain the optimal feature set. The SFFS algorithm is conducted using a linear support vector machine (SVM) over 4-fold cross validation. The optimal CC and optimal MLO feature vectors are then used to train two SVMs independently. The output of each SVM is fused and used to train a final SVM. All SVMs are trained and tested using 5-fold cross validation. Since our dataset is slightly imbalanced, we embed the synthetic minority oversampling technique (SMOTE) into each fold to resample the minority cases to balance the training dataset.

To assess the performance of the mass classification, the likelihood score of a case being malignant is generated for each case in the test set over cross validation and used to create a receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) and the standard deviation is then computed. Delong's test is used to check for statistical significance between the ROC curves of each pseudo color set at an alpha level of 0.05 [167]. Five other metrics are generated from the sum of the confusion matrices generated by the final classifier in the two-stage classification scheme over each cross-validation fold, namely: accuracy, sensitivity, specificity, precision, and F1 score.

5.3. Results

The initial dataset obtained is the same dataset used in our previous study[162]. Initially, there were 1,065 cases containing four FFDM images representing the left and right CC and MLO view. The ipsilateral matching scheme failed to confirm 66 cases, while the bilateral registration scheme failed in 35 cases. This resulted in 964 cases. Of these 964 cases, we were unable to obtain ground truth mass annotations in both images containing the mass for 134 cases, resulting in a final dataset of 830 cases. Of the 830 cases, 310 of the masses are biopsy proven benign and 520 cases are biopsy proven malignant, which corresponds to 3,320 64x64 greyscale ROIs and 1,660 binary segmentation masks.

After 5-fold CV, automated mass segmentation via a Unet achieved a DSC of 0.894 ∓ 0.002 and a Jaccard Index of 0.814 ∓ 0.003. Some examples of the manual ground truth segmentation and the automated segmentation can be seen in **figure 5-2**.

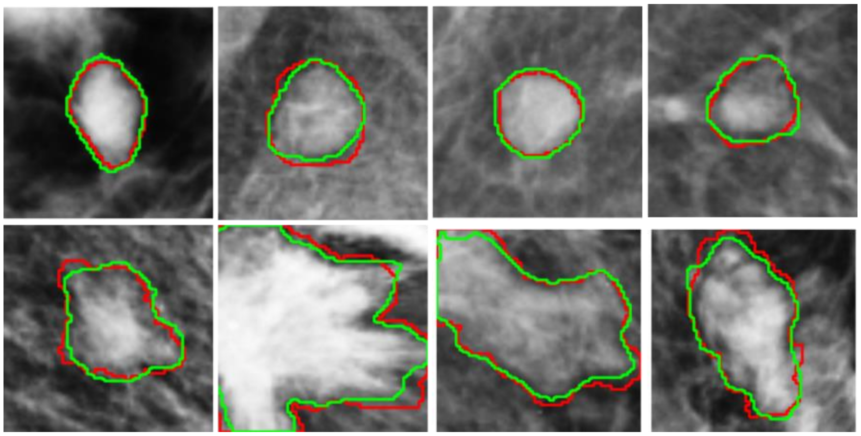

Figure 5-2: *Examples of the manual ground truth segmentation in red and the UNET produced segmentation mask in green. The top row are benign cases, and the bottom row are malignant cases.*

The performance metrics of the mass classification scheme for each pseudo color image set can be seen in **figures 5-3-5-4** and **table 5-3**. Overall, sets B_gt and D_gt outperform all other sets in terms of AUC, accuracy, sensitivity, specificity, precision, and F1 score. Pseudo color sets B_gt and D_gt, both created using the ground truth segmentation masks, perform significantly better than sets A, B_unet, C, D_unet, and E in terms of AUC (p=0.0046, 0.0003, 0.0066, <0.0001, <0.0001 and p=0.0013, 0.0001, 0.0025, 0.0001, <0.0001, respectively). There are no statistically significant differences between the AUC of set B_gt and set D_gt (p=0.6036) or between sets A, B_unet, C, or D_unet. All sets perform significantly better than set E in terms of AUC (p<0.001 for all comparisons).

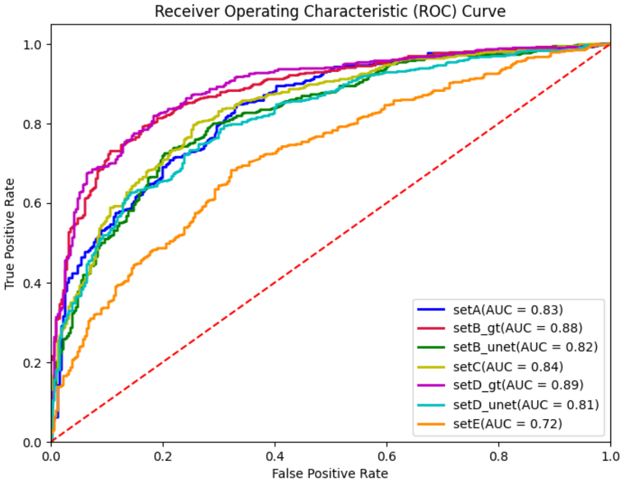

Figure 5-3: *Final receiver operating characteristic curves for all seven pseudo color image sets.*

	AUC	Acc	Precision	Sensitivity	Specificity	F1 score
A	0.833 ± 0.014	0.747 ± 0.030	0.817 ± 0.028	0.771 ± 0.068	0.706 ±0.072	0.791 ± 0.033
B_gt	0.882 ± 0.012	0.812 ± 0.009	0.864 ± 0.016	0.831 ± 0.016	0.781 ± 0.032	0.847 ± 0.008
B_unet	0.820 ± 0.015	0.763 ± 0.022	0.825 ± 0.025	0.790 ± 0.035	0.716 ± 0.053	0.806 ± 0.019
C	0.836 ± 0.014	0.777 ± 0.028	0.833 ± 0.020	0.0806 ± 0.037	0.729 ± 0.034	0.819 ± 0.025
D_gt	0.889 ± 0.012	0.816 ± 0.020	0.871 ± 0.021	0.829 ± 0.025	0.794 ± 0.039	0.849 ± 0.017
D_unet	0.812 ± 0.015	0.741± 0.0250	0.816± 0.0315	0.758 ± 0.017	0.713 ± 0.059	0.786 ± 0.018
E	0.718 ± 0.018	0.681 ± 0.035	0.767 ± 0.030	0.704 ± 0.039	0.642 ± 0.050	0.734 ± 0.031

Table 5-3: *Mean and standard deviation of all metrics computed from each test fold over 5-fold cross validation.*

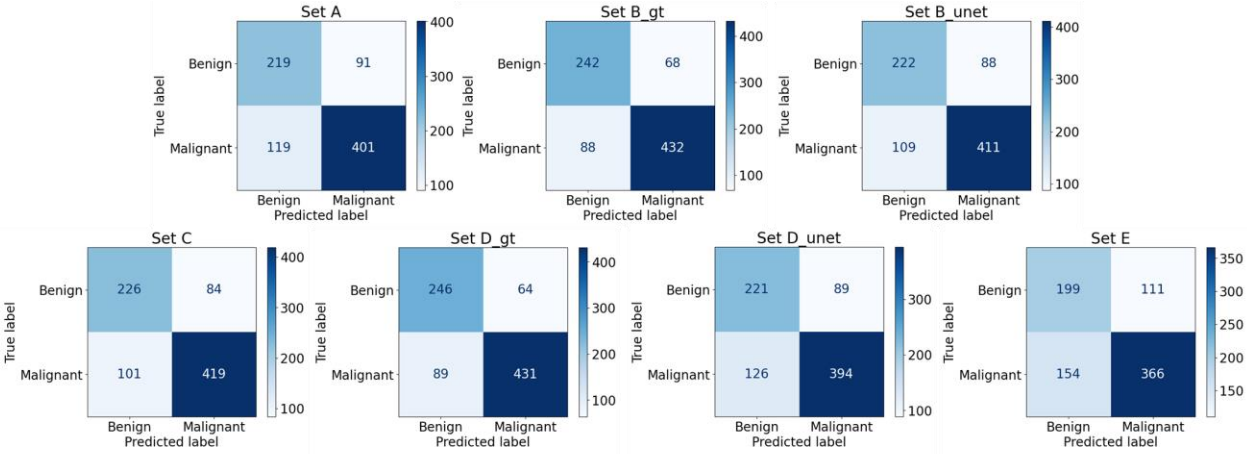

Figure 5-4: *Confusion matrices for each of the seven pseudo color sets. Within each matrix, the top left corner represents the number of true negatives (TN), the top right corner represents the number of false positives (FP), the bottom left corner represents the number of false negatives (FN), and the bottom right corner represents the number of true positives (TP).*

While the performance of set B_gt and set D_gt support our hypothesis that pseudo color images created using a segmentation mask will provide increased information and yield better performance than pseudo color images created without the additional morphological information, the performance of set B_unet and set D_unet is not significantly better than sets A or C. But since the performance of set B_gt and D_gt are significantly better than sets B_unet and D_unet, this can be attributed to the quality of the automated UNET segmentation step as the only difference between these two sets is the creation of the segmentation mask that is in the green channel of each image. To further investigate this, we visually inspected the results of the UNET and observed that the network had trouble capturing spiculations to the same degree that is present in the ground truth images (**Figure 5-5**).

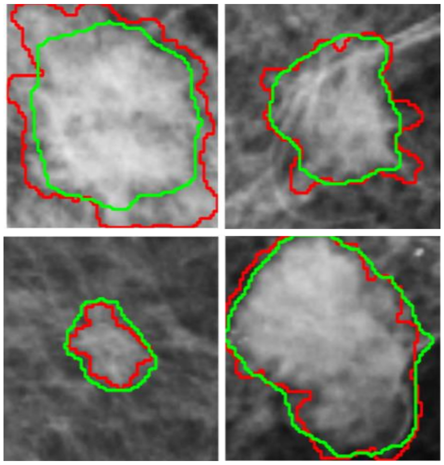

Figure 5-5: *Examples of malignant cases where the UNET does a poor job at capturing spiculations. The ground truth segmentation is in red and the UNET produced segmentation mask is in green.*

Since the morphology of malignant lesions is associated with spiculated margins, the inability to capture spiculations may trick the scheme into believing a lesion is benign since the segmentation mask is rounded when there are actually spiculations present. This would result in an increase in the number of false negatives, which is seen between sets B_gt and B_unet and D_gt and D_unet (**Figure 5-5**). We believe this explains the reason for the decrease in performance when using the automated mass segmentation as opposed to the ground truth in the pseudo color sets.

5.4. Discussion

In this work we investigate the effects of pseudo color image generation on classifying suspicious breast lesions as malignant or benign using deep transfer learning. Our previous work began to investigate the effects of creating pseudo color images using various preprocessing techniques to increase the information passed to the deep network with the goal of increasing the performance of a CADx framework in classifying malignant and benign lesions[154]. In that study, two image sets were created; the first set used the original image stacked in three channels, and the second set used the original image in combination with variants of the original single channel image that either suppressed noise or enhanced contrast. While the performance did increase when using the pseudo color images created with pre-processed variants, this was not always a statistically significant difference. Similarly, the studies conducted by Li et al. and Min et al. for detecting masses using pseudo color generated images suggested that pseudo color images that contain morphological information will improve breast mass detection while pseudo color images that contained texturally enhanced versions did not improve performance[94, 95]. In this work, we aim to see if this follows for mass classification as well.

We hypothesize that creating pseudo color images with additional morphological information will provide increased complementary information to a deep network pre-trained on the ImageNet database, and that this will yield better performance in classifying malignant and benign lesions than when using pseudo color images that do not contain morphological information. Overall, the results of this work support our hypothesis and demonstrate that the addition of the segmented mass to the pseudo color images prior to using deep transfer learning significantly improved the ability of the network to classify malignant and benign lesions when compared to pseudo color images created using only the original image and pre-processed variants that improved contrast or decreased noise.

Since sets B_gt and D_gt perform significantly better than all other sets, we can conclude that the addition of the segmented mass image to the pseudo color image is responsible for the increase in performance. Additionally, we do not observe a significant difference between sets B_gt and D_gt which only vary in the green blue channel (AUC=

0.882 and 0.889, p=0.6036). Set B_gt contains a histogram equalized version of the original image in the blue channel, while set D_gt contains another copy of the original image. This indicates that the addition of the histogram equalized image does not also increase performance and that the increase in performance is solely due to the addition of the segmented mass. There are also no statistically significant differences between the AUC values of sets A, B_unet, C, and D_unet. This indicates that using pre-processing techniques that aim to increase the textural information passed to the deep CNN by convolving a filter with the original image may not actually increase the information as the convolutional layers of the CNN may be able to automatically learn similar features without this addition. Our results support this assertion as there are no significant differences between set A, which contains only the original image, and set C which contains two texturally enhanced versions. On the contrary, the CNN is not able to automatically learn the morphological information that the fully segmented mass channel provided without architectural modifications. This indicates that the pseudo color sets containing a fully segmented mass channel provide increased and complementary information to the network which yields significantly better performance while the pseudo color sets created from texturally enhanced variant channels do not increase the information provided therefore do not have classification performance improvements.

In this study, the morphological information is provided by segmenting the suspicious mass from the background tissue. We use two different techniques to obtain the segmented mass: manual segmentation and automated segmentation via a Unet. The manual segmentation mask is used as the ground truth image for the automated segmentation task. We recognize that acquiring a manual segmentation for every suspicious mass is an extremely time consuming and error prone task. To overcome future issues of obtaining this ground truth segmentation, we use a Unet to demonstrate that a fully automated mass segmentation network can be used in place of a manual segmentation mask. The results of this study show that the pseudo color images created using the Unet generated segmentation mask do not perform as well as the pseudo color images created using the manual segmentation mask. Visual inspection of the Unet generated segmentation masks revealed that the network was doing a poor job at capturing spiculations (**Figure 5-6**) which may be the reason for the decrease in performance. We believe that further work into creating a better automated mass segmentation network will overcome this problem. In this study, a basic Unet architecture is used with only 1,328 examples in the training dataset in each fold. Many complex modifications to the Unet architecture have been proposed in breast mass segmentation tasks that should be investigated in this framework[163, 168, 169]. In addition to adding more robust training data and modifying the Unet architecture, other segmentation networks should be investigated to improve performance, including SegNet, Fully Convolutional Networks (FCN), and conditional generative adversarial networks (cGAN) as these networks have shown superior performance in breast mass segmentation tasks[170-172].

While investigating the best method for the classification portion of the framework, there were extensive attempts to fine tune the VGG16 network as opposed to using it as a feature extractor. We were unable to successfully train a multi-view model for pseudo color sets A and C as there was an overfitting problem that could not be overcome unless we had a larger dataset. This may also support the conclusion that the addition of the morphological information to sets B and D did provide increased information which mitigated the overfitting issue experienced by sets A and C.

While this is a proof-of-concept study, we faced some notable limitations. First, the dataset used in this work is acquired from a single location. Therefore, we are unsure if the results will hold up when using datasets from other locations with mammography images acquired from different machines with different scanning protocols. Second, it is extremely difficult to obtain the ground truth mass segmentation images. In this study alone, there were 134 cases which had to be removed due to the inability to draw the annotations due to dense breast tissue obstructing the view or local irregularities making it difficult to find the boundary. We recognize that this may hinder others from using this technique in future computer aided diagnosis schemes. Using a fully automated segmentation network trained on a large and diverse mammography image set will allow overcome this limitation. Third, we only investigate and compare seven different pseudo color image sets that are made up of combinations of four single channel images, the original image, a bilaterally filtered image, a histogram equalized image, and the segmented mass. While the decision to use a bilaterally filtered image and histogram equalized image was to decrease noise and increase contrast as mammograms are x-ray images which are traditionally noisy and low contrast, there are many different pre-processing techniques that are commonly used in mammography based CADx systems that can be investigated further[173].

5.5. Conclusions

As deep learning techniques continue to outperform traditional machine learning techniques, it is important to experiment with ways in which these networks are used in mammography based CADx schemes. The need for a large and diverse dataset to train a deep CNN often forces researchers to use a transfer learning technique in lieu of training a network from scratch. Utilizing a state-of-the-art deep CNN pretrained on the ImageNet dataset for the breast mass classification task requires some manipulation of the network or input images before training. The results of this study demonstrate that using pseudo color images that include increased morphological information as input to a pre-trained VGG16 network will improve the performance abilities in classifying malignant and benign lesions.

Chapter 6. Conclusions and future work

6.1. Summary

Breast cancer remains an extremely deadly disease with incidence on the rise. Early detection through routine screening exams remains the best method for reducing the mortality associated with the disease. However, the efficacy including both sensitivity and specificity of current breast screening must be improved. The increase in the number of breast imaging modalities coupled with a large amount of clinical, pathological, and genetic information has made it more difficult and time consuming for clinicians to digest all available information and make an accurate diagnosis and appropriate personalized treatment plan. Recent advances in radiomics and DL technology provide promising opportunities to extract more clinically relevant image features as well as to streamline many different types of diagnostic information to build novel CAD systems as decision-making support tools that aim to help clinicians make more accurate and robust cancer diagnosis and treatment decisions.

In summary, the work presented in this dissertation focuses on investigating and developing different methods to improve the performance of CADx systems for mammography by increasing the feature representation of the input images.

In chapter 3 we increase the feature representation of suspicious mammography detected masses by fusing a handcrafted radiomics feature set with a deep transfer learning generated feature set. Our study concluded that the CADx scheme that uses the fusion feature set performs significantly better at classifying masses as malignant or benign than the same scheme using either only handcrafted radiomics or DL features.

While I was not the first person to investigate the fusion of ML and DL-based CAD techniques, the contributions of the work include the following. Firstly, we demonstrate that even though using only DL features outperforms the use of only handcrafted features, the two sets are complementary therefore when fused together performance is significantly improved. This signifies that the domain expertise included in handcrafted feature extraction is useful and should not be ignored. Second, we develop a novel feature selection and reduction pipeline that is able to successfully extract the most meaningful features from an extremely high dimensional feature pool. Third, we began to investigate the effects of pseudo color image generation on the DL feature extraction step. We observed better classification performance when using pseudo color images that contained pre-processed variants, highlighting the importance of preprocessing in mammography-based CAD, and providing us motivation to continue the investigation into pseudo color images in CADx which is done in chapter 5.

In chapter 4 we increase the feature representation of suspicious mammography detected masses by including both the CC and MLO view images of the mass and the contralateral breast and by using a fusion of radiomics and DL features. Our study can be differentiated from the existing multi-view CAD studies in three ways, namely: the inclusion of both radiomics and DL features, the true case-based nature of the input images, and the quantification of the bilateral asymmetry. We include an ipsilateral matching scheme and bilateral registration scheme to ensure that the ROIs that pass simultaneously through the framework correspond to the same mass ipsilaterally and the same region bilaterally. The bilateral registration scheme also allows us to obtain a more accurate quantification of the bilateral asymmetry.

In chapter 5 we increase the feature representation of suspicious mammography detected masses by generating pseudo color images that include increased morphological information. In this study we build off our work in chapter 3 and chapter 4 by continuing to investigate the role of pseudo color image generation in the deep transfer learning feature extraction step of a multi-view CAD system. The results demonstrate that pseudo color image sets that contain increased morphological information perform significantly better than any other set in classifying breast masses as malignant or benign. This work demonstrates the feasibility of improving classification performance when using transfer learning techniques through a relatively simple image transformation. To the best of our knowledge, no other study investigates the link between pseudo color image generation and mammography based CADx performance.

Over the past three years, I have made great progress in understanding the indisputable role that artificial intelligence continues to have in the medical imaging field. I have had the opportunity to investigate and develop new methods for improving the accuracy of mammography-based CAD systems. While the focus of this dissertation is on mammography imaging as it is the most widely used and accessible breast imaging modality, there are other imaging modalities which cannot be ignored. In an effort to give myself a well-rounded understanding of breast imaging beyond mammography, I published a review paper that details current advances in CAD schemes for breast cancer that includes all breast imaging modalities. Additionally, I have published and co-authored multiple journal articles and conference papers.

6.1.1. Journal Papers

1. **Jones, MA**. Faiz, R. Islam, W. Qiu, Y. Pseudo Color Image Generation for Improving the Performance of Deep Transfer Learning-based Computer Aided Diagnosis Schemes in Breast Mass Classification. (submitted to Computers and Biology in Medicine)

2. **Jones, MA**. Sadeghipour, N. Chen, X. Islam, W. and Zheng, B. A multi-stage fusion framework to classify breast lesions using deep learning and radiomics features computed from four-view mammograms. *Med Phys.* 2023 March 31.

3. Sheth, V. Chen, X. Mettenbrink, EM. Yang, W. **Jones, MA**. M'Saad, O. Thomas, A. Newport, RS. Francek, E. Wang, L. Frickenstein, AN. Donahue, N. Holden, A. Mjema, NF. Green, DE, DeAngelis, PL. Bewersdorf, J. Wilhem, S. Quantifying Intracellular Nanoparticle Distributions with Three-Dimensional Super-Resolution Microscopy. *ACS Nano*. 2023 April 18.

4. Islam, W. **Jones, MA**. Faiz, R. Sadeghipour, N. Qiu, Y. Zheng, B. Improving Performance of Breast Lesion Classification Using a RestNet50 Model Optimized with a Novel Attention Mechanism. *Tomography.* 2022 Sept 28.

5. **Jones, MA**. Islam, W. Faiz, R. Chen, X. and Zheng, B. Applying artificial intelligence technology to assist with breast cancer diagnosis and prognosis prediction. *Front. Oncol.* 2022 Aug.

6. Danala, G. Maryada, S.K. Islam, W; Faiz, R. **Jones, MA**. Qiu, Y; Zheng, B. A comparison of Computer-Aided Diagnosis Schemes Optimized Using

Radiomics and Deep Transfer Learning Methods. *Bioengineering*. 2022 June 13.

7. **Jones, MA.** Faiz, R. Qiu, Y. Zheng, B. Improving Mammography Lesion Classification by Optimal Fusion of Handcrafted and Deep Transfer Learning Features. *Physics in Medicine & Biology*. 2022 Feb 7.

8. Gai, T. Thai, T. **Jones, MA.** Jo, J. Zheng, B. Applying a radiomics-based CAD scheme to classify between malignant and benign pancreatic tumors using CT images. *J Xray Sci Technol*. 2022 Jan 24.

9. **Jones, MA.** MacCuaig, WM. Frickenstein, AN. Camalan, S. Gurcan, M.N. Holter-Chakrabarty, J. Morris, K.T. McNally, M.W. Booth, K.K. Carter, S. Grizzle, W.E. McNally, L.R. Molecular Imaging of Inflammatory Disease. *Biomedicines* 2021 Feb 4.

10. MacCuaig, WM*. **Jones, MA*.** Abeyakoon, O. McNally, LR. Development of Multispectral Optoacoustic Tomography (MSOT) as a clinically translatable imaging modality. *Radiology: Imaging Cancer* 2020 Nov 20. **(*Co-First authors)**

11. Frickenstein, A*. **Jones, MA***. Behkam, B. McNally, LR. Imaging inflammation and infection in the gastrointestinal tract. *Int J Mol Sci*. 2019 Dec 30. **(*Co-First authors)**

12. Gomez-Gutierrez, JG*. Bhutiani, N*. McNally, MW. Chuong, P. Yin, W. **Jones, MA.** Zeiderman, MR. Grizzle, WE. McNally, LR. The neutral red assay can be

used to evaluate cell viability during autophagy or in an acidic microenvironment in vitro. *Biotechnic and Histochemistry.* 2020 Aug 03.

6.1.2. Conference Papers

1. Sadeghipour, N. Tabesh, F. Natarajan, A. **Jones, MA**. Chen, X. Paulmurugan, R. and Zheng,B. Quantitative methods for molecular ultrasound imaging. SPIE Medical Imaging 2023. April 10th, 2023.

2. **Jones MA**, Pham H, Gai T, Zheng B. Fusion of Handcrafted and Deep Transfer Learning Features to Improve Performance of Breast Lesion Classification. Stephenson Cancer Center 2022 Cancer Research Symposium. Oklahoma City, Oklahoma. March 4th, 2022

3. **Jones MA**, Pham H, Gai T, Zheng B. Fusion of Handcrafted and Deep Transfer Learning Features to Improve Performance of Breast Lesion Classification. SPIE Medical Imaging 2022. San Diego, CA. February 2022.

4. Danala G, Mirniaharikandehei S, **Jones MA**, Gai T, Maryada SK, Wu D, Qiu Y, Zheng B. Developing interactive computer-aided detection tools to support translational clinical research. Proc. SPIE 12035, Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment. March 21st 2022.

5. Mirniaharikandehei S, Hollingsworth A, **Jones MA**, Liu H, Qiu Y, Zheng B. Assessment of a new CAD-generated imaging marker to predict risk of having

mammography-occult tumors. SPIE Medical Imaging. San Diego, CA. February 2022.

6. Pham H, **Jones MA**, Gai T, Islam W, Danala G, Jo J, Zheng B. Identifying an optimal machine learning generated image marker to predict survival of gastric cancer patients. SPIE Medical Imaging. San Diego, CA. February 2022.

7. Danala G, Maryada SK, Pham H, Islam W, **Jones MA**, Zheng B. Comparison of performance in breast lesions classification using radiomics and deep transfer learning: An assessment study. SPIE Medical Imaging. San Diego, CA. February 2022.

8. **Jones MA**, Fouts B, McNally M, Samkutty A, MacCuaig W, Frickenstein AN, McNally LR. Evaluation of multispectral separation algorithms to identify spectrally distinct chromophores in breast cancer. AACR Annual Meeting 2020. San Diego, CA. April 2020.

9. Frickenstein A**,** MacCuaig W, **Jones MA**, Fouts B, Samkutty A, McNally M, McNally LR. Targeting behavior and pharmacokinetics of pHILP-conjugated mesoporous silica nanoparticles in pancreatic tumors. AACR Annual Meeting 2020. San Diego, CA. April 2020.

10. **Jones MA**, Fouts B, McNally M, Samkutty A, McNally LR. Breast cancer treatment by pH responsive mesoporous silica nanoparticles. Stephenson

Cancer Center 2020 Cancer Research Symposium. Oklahoma City, Oklahoma. February 7th, 2020.

11. **Jones MA**, Fouts B, McNally M, Samkutty A, McNally LR. Development of pH responsive mesoporous silica nanoparticles in the treatment of ER/PR + Breast Cancer. END2Cancer, Oklahoma City, Oklahoma. November 2019

6.2. Future work

Despite the extensive research that has been conducted in developing CAD schemes to aid radiologists in reading and interpretating mammography images, there are still many challenges that must be addressed for these systems to be robust enough to proceed to clinical use. There are many generic challenges that almost all CAD systems face which have been discussed in the introduction (section 1.3) of this dissertation. Notable challenges and future goals specific to the research presented in this dissertation are as follows.

First, the mammography images used in chapter 3, 4, and 5 all come from the same dataset. It is unknown how well these studies will generalize on unseen datasets that come from different locations and different scanners. This highlights the important obstacle that is the lack of large and high-quality image databases for many different application tasks. Although several breast image databases are publicly available including DDSM, INbreast, MIAS, and BCDR, these databases mainly contain easy cases and lack subtle cases, which substantially reduces the diversity and heterogeneity of these image databases. Many existing databases reported in previous research papers are also either obsolete (i.e., DDSM and MIAS used the digitized screen-film based mammograms) or have a lack of biopsy-approved ground-truth (i.e., INbreast). Thus, models developed using these “easy” databases have lower performance when applied to real diverse images acquired in clinical practice. In our work, we use a private mammography database that more accurately depicts real-life clinical data as it contains a diverse imaging set with high heterogeneity in the lesions as well as breast densities.

By recognizing such limitations or challenges, more research efforts continue to build better public image databases. For example, The Cancer Imaging Archive (TCIA) was created in 2011 with the aim of developing a large, de-identified, open-access archive of medical images from a wide variety of cancers and imaging modalities [174]. New significant progress is expected in future studies to build this important infrastructure to help develop robust predictive models in the medical imaging field. Thus, the establishment of TCIA allows researchers to train and validate their prediction models on imaging data acquired from other clinical sites to help researchers develop more accurate and robust models that can eventually be translated to the clinic.

Second, our work focuses solely on mammography imaging techniques despite there being other breast cancer imaging modalities. The downfalls of mammography have led to an increase in the use of other adjunct imaging modalities in clinical practice including ultrasound (US) and dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) [175, 176]. Digital breast tomosynthesis (DBT) is a newer modality that is commonly used in which X-ray images are taken over multiple angles in a limited range (i.e., $\pm 15^{\circ}$) and the acquired scanning data is reconstructed into quasi-3D breast images to reduce the impact of dense breast tissue overlap in 2D mammograms [177]. Additionally, several other new imaging modalities including contrast enhanced spectral mammography (CESM) [175, 176], phase contrast breast imaging [178], breast computed tomography [179], thermography and electrical impedance tomography of breast imaging [180], and molecular breast imaging [181], have also been investigated and tested in many prospective studies or clinical trials. The frameworks of the studies described in this dissertation should be applied to other imaging modalities when applicable. Mammography remains the most commonly used and accessible breast imaging modality worldwide, therefore the work conducted in this dissertation is still extremely relevant.

Third, in our studies a VGG16 network pretrained on the ImageNet database is used as a feature extractor. The decision is rooted in its established success and widespread applicability. However, it is imperative to acknowledge that the landscape of deep learning for medical image classification tasks is continually evolving, offering a multitude of state-of-the-art alternatives. The alternative networks should be investigated further, not only as feature extractor but also through fine tuning.

Fourth, we focus this dissertation on the improvement of classifying malignant and benign breast masses. I believe that the trajectory of this work should be expanded to include a fully automated detection step prior to mass classification. Currently, the mass detection step is conducted manually as a radiologist has marked the center of each suspicious region which is used as a guide to extract ROIs. Many existing studies focus solely on the detection of suspicious regions rather than the classification. We believe that future work should be conducted to create a fully automated mass detection and classification system. Such modification would enhance the clinical applicability, ultimately paving the way for the integration of these systems into clinical practice.

Fifth, it is important to note that we reported our overall AUC values at a threshold of 0.5 as this represents a balanced trade-off between the true positive rate and true negative rate, but this metric may not be ideal in the clinical context of classifying suspicious lesions as malignant or benign[182]. Clinically, it is most important to limit the number of false negatives, as this would mean that an individual that does have breast cancer is told that she does not have cancer which has significant consequences as early treatment gives the best chance at survival. In order to ensure that this false negative rate is low, we can decrease the threshold of our model which will result in an increase in sensitivity as more cases are being predicted as positive but a decrease in specificity as this increases the number of false positives. This trade-off results in more women with benign lesions undergoing further testing but this is considered a clinically acceptable outcome compared to letting malignant lesions go undiagnosed. Before this work can be translated to the clinic, the choice of threshold must be carefully investigated with the help of experienced breast radiologists. We will look at how the sensitivity and specificity of our model change with varied thresholds. This domain expertise will allow us to choose an optimal threshold that minimizes the false positive rate while also ensuring that the sensitivity and specificity are acceptable.

Lastly, a graphical user interface should be created which can be used and tested in clinical practice. The performance of AI-based models reported in the literature based on laboratory studies may not be directly applicable to clinical practice as researchers have found that higher sensitivity of experimental CAD systems may not actually help radiologists in reading and interpreting images in clinical practice. One previous observer performance study reported that radiologists failed to recognize correct prompts of CADe scheme in 71% of missed cancer cases due to higher false-positive prompts [158]. By retrospectively analyzing a large cohort of clinical data before and after implementing CADe schemes in multiple community hospitals, one study reported that the current method of using CADe schemes in mammography reduced radiologists' performance as seen by decreased specificity and positive predictive values [183]. In order to overcome this issue, researchers have investigated several new approaches of using CADe schemes. One study reported that using an interactive prompt method to replace a conventional "second reader" prompt method significantly improves radiologists' performance in detecting malignant masses from mammograms [159]. However, this interactive prompting method has not been accepted in clinical practice. Thus, the lessons learned from CADe schemes used in clinical practice indicate that more research efforts are needed to investigate and develop new methods, including FDA clearance processes, to evaluate the potential clinical utility of all new CAD systems for different clinical medical imaging applications [184].

Chapter 7. References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA: a cancer journal for clinicians,* https://doi.org/10.3322/caac.21708 vol. 72, no. 1, pp. 7-33, 2022/01/01 2022, doi: https://doi.org/10.3322/caac.21708.

[2] C. E. DeSantis *et al.*, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians,* https://doi.org/10.3322/caac.21583 vol. 69, no. 6, pp. 438-451, 2019/11/01 2019, doi: https://doi.org/10.3322/caac.21583.

[3] L. Berlin and F. M. Hall, "More mammography muddle: emotions, politics, science, costs, and polarization," *Radiology,* vol. 255, no. 2, pp. 311-316, 2010.

[4] J. McCann, D. Stockton, and S. Godward, "Impact of false-positive mammography on subsequent screening attendance and risk of cancer," *Breast Cancer Research,* vol. 4, no. 5, pp. 1-9, 2002.

[5] P. C. Gøtzsche, "Mammography screening is harmful and should be abandoned," (in eng), *J R Soc Med,* vol. 108, no. 9, pp. 341-5, Sep 2015, doi: 10.1177/0141076815602452.

[6] M. Brennan and N. Houssami, "Discussing the benefits and harms of screening mammography," *Maturitas,* vol. 92, pp. 150-153, 2016/10/01/ 2016, doi: https://doi.org/10.1016/j.maturitas.2016.08.003.

[7] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," (in eng), *Br J Radiol,* vol. 95, no. 1130, pp. 20211033-20211033, 2022, doi: 10.1259/bjr.20211033.

[8] A. Yala *et al.*, "Toward robust mammography-based models for breast cancer risk," (in eng), *Sci Transl Med,* vol. 13, no. 578, Jan 27 2021, doi: 10.1126/scitranslmed.aba4373.

[9] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Computers in Biology and Medicine,* p. 106554, 2023.

[10] J. Katzen and K. Dodelzon, "A review of computer aided detection in mammography," (in eng), *Clinical imaging,* vol. 52, pp. 305-309, Nov-Dec 2018, doi: 10.1016/j.clinimag.2018.08.014.

[11] M. D. Dorrius, M. C. der Weide, P. van Ooijen, R. M. Pijnappel, and M. Oudkerk, "Computer-aided detection in breast MRI: a systematic review and meta-analysis," *European radiology,* vol. 21, no. 8, pp. 1600-1608, 2011.

[12] F. Winsberg, M. Elkin, J. Macy Jr, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis," *Radiology,* vol. 89, no. 2, pp. 211-215, 1967.

[13] W. Spiesberger, "Mammogram inspection by computer," *IEEE Transactions on Biomedical Engineering,* no. 4, pp. 213-219, 1979.

[14] S. A. Feig and M. J. Yaffe, "Digital mammography, computer-aided diagnosis, and telemammography," *Radiologic Clinics of North America,* vol. 33, no. 6, pp. 1205-1230, 1995.

[15] H.-P. Chan *et al.*, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," *Investigative radiology,* vol. 25, no. 10, pp. 1102-1110, 1990.

[16] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology,* vol. 220, no. 3, pp. 781-786, 2001.

[17] J. D. Keen, J. M. Keen, and J. E. Keen, "Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016," *Journal of the American College of Radiology,* vol. 15, no. 1, pp. 44-48, 2018.

[18] A. Rodríguez-Ruiz *et al.*, "Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System," *Radiology,* vol. 290, no. 2, pp. 305-314, 2019/02/01 2018, doi: 10.1148/radiol.2018181371.

[19] J. J. Fenton *et al.*, "Influence of Computer-Aided Detection on Performance of Screening Mammography," *New England Journal of Medicine,* vol. 356, no. 14, pp. 1399-1409, 2007, doi: 10.1056/NEJMoa066099.

[20] E. L. Henriksen, J. F. Carlsen, I. M. Vejborg, M. B. Nielsen, and C. A. Lauridsen, "The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review," (in eng), *Acta Radiol,* vol. 60, no. 1, pp. 13-18, Jan 2019, doi: 10.1177/0284185118770917.

[21] "Food and drug administration. cmTriage," ed, 2019.

[22] "Food and drug administration. HealthMammo," ed, 2020.

[23] "Food and drug administration. Saige-Q," ed, 2021.

[24] "Food and drug administration. MammoScreen," ed, 2020.

[25] "Food and drug administration. Genius AI detection," ed, 2020.

[26] "Food and drug administration. ProFound AI software V2.1," ed, 2019.

[27] "Food and drug administration. 510(k) premarket notification," ed, 2021.

[28] "Food and drug administration. Transpara 1.7.0," ed, 2021.

[29] "Food and drug administration. Lunit INSIGHT MMG," ed, 2021.

[30] R. Ranjbarzadeh *et al.*, "Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods," (in eng), *Comput Biol Med,* vol. 152, p. 106443, Jan 2023, doi: 10.1016/j.compbiomed.2022.106443.

[31] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, "Breast cancer segmentation methods: current status and future potentials," *BioMed Research International,* vol. 2021, pp. 1-29, 2021.

[32] Z. Rezaei, "A review on image-based approaches for breast cancer detection, segmentation, and classification," *Expert Systems with Applications,* vol. 182, p. 115204, 2021/11/15/ 2021, doi: https://doi.org/10.1016/j.eswa.2021.115204.

[33] M. Caballo *et al.*, "Computer-aided diagnosis of masses in breast computed tomography imaging: deep learning model with combined handcrafted and convolutional radiomic features," (in eng), *J Med Imaging (Bellingham),* vol. 8, no. 2, p. 024501, Mar 2021, doi: 10.1117/1.Jmi.8.2.024501.

[34] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," *Alexandria Engineering Journal,* vol. 60, no. 5, pp. 4701-4709, 2021/10/01/ 2021, doi: https://doi.org/10.1016/j.aej.2021.03.048.

[35] M. Tan, W. Qian, J. Pu, H. Liu, and B. Zheng, "A new approach to develop computer-aided detection schemes of digital mammograms," (in eng), *Phys Med Biol,* vol. 60, no. 11, pp. 4413-27, Jun 7 2015, doi: 10.1088/0031-9155/60/11/4413.

[36] M. Heidari, S. Mirniaharikandehei, G. Danala, Y. Qiu, and B. Zheng, "A new case-based CAD scheme using a hierarchical SSIM feature extraction method to classify between malignant and benign cases," in *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, 2020, vol. 11318: International Society for Optics and Photonics, p. 1131816.

[37] H. Li *et al.*, "Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location," (in eng), *Med Phys,* vol. 31, no. 3, pp. 549-55, Mar 2004, doi: 10.1118/1.1644514.

[38] S. Rizzo *et al.*, "Radiomics: the facts and the challenges of image analysis," *European radiology experimental,* vol. 2, no. 1, pp. 1-8, 2018.

[39] P. Lambin *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature reviews Clinical oncology,* vol. 14, no. 12, pp. 749-762, 2017.

[40] M. Goto *et al.*, "Diagnosis of breast tumors by contrast-enhanced MR imaging: comparison between the diagnostic performance of dynamic enhancement patterns and morphologic features," (in eng), *J Magn Reson Imaging,* vol. 25, no. 1, pp. 104-12, Jan 2007, doi: 10.1002/jmri.20812.

[41] Y. Cui, Y. Li, D. Xing, T. Bai, J. Dong, and J. Zhu, "Improving the Prediction of Benign or Malignant Breast Masses Using a

Combination of Image Biomarkers and Clinical Parameters," (in eng), *Front Oncol,* vol. 11, pp. 629321-629321, 2021, doi: 10.3389/fonc.2021.629321.

[42] C. Varela, S. Timp, and N. Karssemeijer, "Use of border information in the classification of mammographic masses," (in eng), *Phys Med Biol,* vol. 51, no. 2, pp. 425-41, Jan 21 2006, doi: 10.1088/0031-9155/51/2/016.

[43] T. Wang, J. Gong, H. H. Duan, L. J. Wang, X. D. Ye, and S. D. Nie, "Correlation between CT based radiomics features and gene expression data in non-small cell lung cancer," (in eng), *Journal of X-ray science and technology,* vol. 27, no. 5, pp. 773-803, 2019, doi: 10.3233/xst-190526.

[44] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics,* no. 6, pp. 610-621, 1973.

[45] W. H. Nailon, "Texture analysis methods for medical image characterisation," *Biomedical imaging,* vol. 75, p. 100, 2010.

[46] R. E. Bellman, *Adaptive Control Process*. Princeton University Press, 1961.

[47] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, no. Mar, pp. 1157-1182, 2003.

[48] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 25-29 May 2015 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

[49] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, no. 2, pp. 228-233, 2001.

[50] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," (in eng), *Comput Methods Programs Biomed,* vol. 156, pp. 25-45, Mar 2018, doi: 10.1016/j.cmpb.2017.12.012.

[51] R. Jalloul, H. K. Chethan, and R. Alkhatib, "A Review of Machine Learning Techniques for the Classification and Detection of Breast Cancer from Medical Images," (in eng), *Diagnostics (Basel),* vol. 13, no. 14, Jul 24 2023, doi: 10.3390/diagnostics13142460.

[52] H.-P. Chan, R. K. Samala, and L. M. Hadjiiski, "CAD and AI for breast cancer—recent development and challenges," *Br J Radiol,* vol. 93, no. 1108, p. 20190580, 2019.

[53] Q. Zheng, M. Yang, X. Tian, X. Wang, and D. Wang, "Rethinking the Role of Activation Functions in Deep Convolutional Neural Networks for Image Classification," *Engineering Letters,* vol. 28, no. 1, 2020.

[54] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," *Neural Computation,* vol. 29, no. 9, pp. 2352-2449, 2017, doi: 10.1162/neco_a_00990.

[55] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3D deep learning on medical images: a review," *Sensors,* vol. 20, no. 18, p. 5097, 2020.

[56] F. Gao *et al.*, "SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis," *Computerized Medical Imaging and Graphics,* vol. 70, pp. 53-62, 2018/12/01/ 2018, doi: https://doi.org/10.1016/j.compmedimag.2018.09.004.

[57] A. H. Yurttakal, H. Erbay, T. İkizceli, and S. Karaçavuş, "Detection of breast cancer via deep convolution neural networks using MRI images," *Multimedia Tools and Applications,* vol. 79, no. 21, pp. 15555-15573, 2020/06/01 2020, doi: 10.1007/s11042-019-7479-6.

[58] Y. Qiu *et al.*, "A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology," (in eng), *Journal of X-ray science and technology,* vol. 25, no. 5, pp. 751-763, 2017, doi: 10.3233/XST-16226.

[59] L. Alzubaidi *et al.*, "Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study," *Applied Sciences,* vol. 10, no. 13, p. 4523, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/13/4523.

[60] H. C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," (in eng), *IEEE Trans Med Imaging,* vol. 35, no. 5, pp. 1285-98, May 2016, doi: 10.1109/tmi.2016.2528162.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009: Ieee, pp. 248-255.

[62] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *arXiv preprint arXiv:1902.07208,* 2019.

[63] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence,* vol. 9, no. 2, pp. 85-112, 2020.

[64] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *International Journal of Automation and Computing,* vol. 14, no. 2, pp. 119-135, 2017.

[65] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging,* vol. 22, no. 1, pp. 1-13, 2022.

[66] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," (in eng), *Comput Methods Programs Biomed,* vol. 127, pp. 248-57, Apr 2016, doi: 10.1016/j.cmpb.2015.12.014.

[67] M. A. Jones, W. Islam, R. Faiz, X. Chen, and B. Zheng, "Applying artificial intelligence technology to assist with breast cancer diagnosis and prognosis prediction," (in English), *Front Oncol,* Review vol. 12, 2022-August-31 2022, doi: 10.3389/fonc.2022.980793.

[68] J. H. Thrall *et al.*, "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," (in eng), *J Am Coll Radiol,* vol. 15, no. 3 Pt B, pp. 504-508, Mar 2018, doi: 10.1016/j.jacr.2017.12.026.

[69] X. T. Li and R. Y. Huang, "Standardization of imaging methods for machine learning in neuro-oncology," (in eng), *Neurooncol Adv,* vol. 2, no. Suppl 4, pp. iv49-iv55, Dec 2020, doi: 10.1093/noajnl/vdaa054.

[70] E. Sala *et al.*, "Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging," (in eng), *Clin Radiol,* vol. 72, no. 1, pp. 3-10, Jan 2017, doi: 10.1016/j.crad.2016.09.013.

[71] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence,* vol. 3, no. 3, pp. 199-217, 2021.

[72] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine,* vol. 17, no. 1, p. 195, 2019/10/29 2019, doi: 10.1186/s12916-019-1426-2.

[73] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," (in eng), *Med Image Anal,* vol. 79, p. 102470, Jul 2022, doi: 10.1016/j.media.2022.102470.

[74] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy,* vol. 23, no. 1, p. 18, 2020.

[75] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923,* 2017.

[76] R. M. Nishikawa and D. Gur, "CADe for early detection of breast cancer—current status and why we need to continue to explore new approaches," *Acad Radiol,* vol. 21, no. 10, pp. 1320-1321, 2014.

[77] S. A. Khan and S.-P. Yong, "A comparison of deep learning and hand crafted features in medical image modality classification," *2016 3rd International Conference on Computer and Information Sciences (ICCOINS),* pp. 633-638, 2016.

[78] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, "Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment," *Scientific Reports,* vol. 10, no. 1, p. 20336, 2020/11/23 2020, doi: 10.1038/s41598-020-77264-y.

[79] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," (in eng), *Med Phys,* vol. 44, no. 10, pp. 5162-5171, Oct 2017, doi: 10.1002/mp.12453.

[80] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," (in eng), *J Med Imaging (Bellingham),* vol. 3, no. 3, pp. 034501-034501, 2016, doi: 10.1117/1.JMI.3.3.034501.

[81] Y. Guo, R. Sivaramakrishna, C.-C. Lu, J. S. Suri, and S. Laxminarayan, "Breast image registration techniques: a survey," *Medical and Biological Engineering and Computing,* vol. 44, no. 1, pp. 15-26, 2006.

[82] B. Zheng *et al.*, "Multiview-based computer-aided detection scheme for breast masses," (in eng), *Med Phys,* vol. 33, no. 9, pp. 3135-43, Sep 2006, doi: 10.1118/1.2237476.

[83] H. Li, K. R. Mendel, L. Lan, D. Sheth, and M. L. Giger, "Digital mammography in breast cancer: additive value of radiomics of breast parenchyma," *Radiology,* vol. 291, no. 1, pp. 15-20, 2019.

[84] H. N. Khan, A. R. Shahid, B. Raza, A. H. Dar, and H. Alquhayz, "Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network," *IEEE Access,* vol. 7, pp. 165724-165733, 2019, doi: 10.1109/ACCESS.2019.2953318.

[85] I. Hina, S. A. Raza, R. Basit, and K. Hasan, "Multi-View Attention-based Late Fusion (MVALF) CADx system for breast cancer using deep learning," *Machine Graphics and Vision,* vol. 29, no. 1/4, pp. 55-78, 12/21 2020, doi: 10.22630/MGV.2020.29.1.4.

[86] K. J. Geras *et al.*, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *arXiv preprint arXiv:1703.07047,* 2017.

[87] S. S. Boudouh and M. Bouakkaz, "Breast cancer: new mammography dual-view classification approach based on pre-processing and transfer learning techniques," *Multimedia Tools and Applications,* 2023/08/14 2023, doi: 10.1007/s11042-023-16431-5.

[88] A. Jouirou, A. Baâzaoui, and W. Barhoumi, "Multi-view information fusion in mammograms: A comprehensive overview," *Information Fusion,* vol. 52, pp. 308-321, 2019.

[89] M. Tan, B. Zheng, P. Ramalingam, and D. Gur, "Prediction of Near-term Breast Cancer Risk Based on Bilateral Mammographic Feature Asymmetry," *Acad Radiol,* vol. 20, no. 12, pp. 1542-1550, 2013/12/01/ 2013, doi: https://doi.org/10.1016/j.acra.2013.08.020.

[90] Y. Li, M. Fan, H. Cheng, P. Zhang, B. Zheng, and L. Li, "Assessment of global and local region-based bilateral mammographic feature asymmetry to predict short-term breast cancer risk," (in eng), *Phys Med Biol,* vol. 63, no. 2, p. 025004, Jan 9 2018, doi: 10.1088/1361-6560/aaa096.

[91] Q. Yang, L. Li, J. Zhang, G. Shao, C. Zhang, and B. Zheng, "Computer-Aided Diagnosis of Breast DCE-MRI Images Using Bilateral Asymmetry of Contrast Enhancement Between Two Breasts," *Journal of Digital Imaging,* vol. 27, no. 1, pp. 152-160, 2014/02/01 2014, doi: 10.1007/s10278-013-9617-4.

[92] N. F. Razali, I. S. Isa, S. N. Sulaiman, N. K. A. Karim, and M. K. Osman, "Color-assisted Multi-input Convolutional Neural Network for Cancer Classification on Mammogram Images," in *2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, 2023: IEEE, pp. 30-35.

[93] P. Teare, M. Fishman, O. Benzaquen, E. Toledano, and E. Elnekave, "Malignancy detection on mammography using dual deep

convolutional neural networks and genetically discovered false color input enhancement," *Journal of digital imaging,* vol. 30, pp. 499-505, 2017.

[94] Y. Li, H. Chen, L. Zhang, and L. Cheng, "Mammographic mass detection based on convolution neural network," in *2018 24th International conference on pattern recognition (ICPR)*, 2018: IEEE, pp. 3850-3855.

[95] H. Min *et al.*, "Fully automatic computer-aided mass detection and segmentation via pseudo-color mammograms and mask r-cnn," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020: IEEE, pp. 1111-1115.

[96] J. Ball, T. Butler, and L. Bruce, "Towards automated segmentation and classification of masses in mammograms," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2004, vol. 1: IEEE, pp. 1814-1817.

[97] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," (in eng), *CA: a cancer journal for clinicians,* vol. 70, no. 1, pp. 7-30, Jan 2020, doi: 10.3322/caac.21590.

[98] A. Bleyer, C. Baines, and A. B. Miller, "Impact of screening mammography on breast cancer mortality," (in eng), *International journal of cancer,* vol. 138, no. 8, pp. 2003-12, Apr 15 2016, doi: 10.1002/ijc.29925.

[99] A. Bleyer and H. G. Welch, "Effect of Three Decades of Screening Mammography on Breast-Cancer Incidence," *New England Journal of Medicine,* vol. 367, no. 21, pp. 1998-2005, 2012/11/22 2012, doi: 10.1056/NEJMoa1206809.

[100] "National Breast Cancer Foundation." https://www.nationalbreastcancer.org/breast-cancer-biopsy. (accessed.

[101] J. Brodersen and V. D. Siersma, "Long-term psychosocial consequences of false-positive screening mammography," (in eng), *Annals of family medicine,* vol. 11, no. 2, pp. 106-15, Mar-Apr 2013, doi: 10.1370/afm.1466.

[102] R. A. Castellino, "Computer aided detection (CAD): an overview," (in eng), *Cancer Imaging,* vol. 5, no. 1, pp. 17-19, 2005, doi: 10.1102/1470-7330.2005.0018.

[103] H. R. Roth *et al.*, "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1170-1181, 2016, doi: 10.1109/TMI.2015.2482920.

[104] M. Tan, J. Pu, and B. Zheng, "Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme," (in eng), *Phys Med Biol,* vol. 59, no. 15, pp. 4357-4373, 2014, doi: 10.1088/0031-9155/59/15/4357.

[105] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014: Springer, pp. 818-833.

[106] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *arXiv preprint arXiv:1810.03292,* 2018.

[107] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 2018: IEEE, pp. 80-89.

[108] R. Paul *et al.*, "Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma," (in eng), *Tomography,* vol. 2, no. 4, pp. 388-395, 2016, doi: 10.18383/j.tom.2016.00211.

[109] R. Paul, S. H. Hawkins, L. O. Hall, D. B. Goldgof, and R. J. Gillies, "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016: IEEE, pp. 002570-002575.

[110] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine,* vol. 8, no. 11, p. 713, 2020. [Online]. Available: https://atm.amegroups.com/article/view/36944.

[111] X. Chen *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *arXiv preprint arXiv:2105.13381,* 2021.

[112] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A Technical Review of Convolutional Neural Network-Based Mammographic Breast Cancer Diagnosis," *Computational and Mathematical Methods in Medicine,* vol. 2019, p. 6509357, 2019/03/25 2019, doi: 10.1155/2019/6509357.

[113] M. Heidari, S. Mirniaharikandehei, W. Liu, A. B. Hollingsworth, H. Liu, and B. Zheng, "Development and Assessment of a New Global Mammographic Image Feature Analysis Scheme to Predict Likelihood of Malignant Cases," (in eng), *IEEE transactions on medical imaging,* vol. 39, no. 4, pp. 1235-1244, 2020, doi: 10.1109/TMI.2019.2946490.

[114] M. Tan, F. Aghaei, Y. Wang, and B. Zheng, "Developing a new case based computer-aided detection scheme and an adaptive cueing method to improve performance in detecting mammographic lesions," (in eng), *Phys Med Biol,* vol. 62, no. 2, pp. 358-376, Jan 21 2017, doi: 10.1088/1361-6560/aa5081.

[115] S. Mirniaharikandehei, A. B. Hollingsworth, B. Patel, M. Heidari, H. Liu, and B. Zheng, "Applying a new computer-aided detection scheme generated imaging marker to predict short-term breast cancer risk," (in eng), *Phys Med Biol,* vol. 63, no. 10, p. 105005, May 15 2018, doi: 10.1088/1361-6560/aabefe.

[116] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu, and B. Zheng, "Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms," *International Journal of Medical Informatics,* vol. 144, p. 104284, 2020/12/01/ 2020, doi: https://doi.org/10.1016/j.ijmedinf.2020.104284.

[117] S. Wu, S. Yu, Y. Yang, and Y. Xie, "Feature and Contrast Enhancement of Mammographic Image Based on Multiscale Analysis and Morphology," *Computational and Mathematical Methods in Medicine,* vol. 2013, p. 716948, 2013/12/12 2013, doi: 10.1155/2013/716948.

[118] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "A gentle introduction to bilateral filtering and its applications," presented at the ACM SIGGRAPH 2007 courses, San Diego, California, 2007. [Online]. Available: https://doi.org/10.1145/1281500.1281602.

[119] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[120] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," *Scientific Reports,* vol. 9, no. 1, p. 12495, 2019/08/29 2019, doi: 10.1038/s41598-019-48995-4.

[121] S. Montaha *et al.*, "BreastNet18: A High Accuracy Fine-Tuned VGG16 Model Evaluated Using Ablation Study for Diagnosing Breast Cancer from Enhanced Mammography Images," *Biology,* vol. 10, no. 12, p. 1347, 2021. [Online]. Available: https://www.mdpi.com/2079-7737/10/12/1347.

[122] S. J. S. Gardezi, M. Awais, I. Faye, and F. Meriaudeau, "Mammogram classification using deep learning features," in *2017 IEEE*

International Conference on Signal and Image Processing Applications (ICSIPA), 2017: IEEE, pp. 485-488.

[123] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access,* vol. 9, pp. 71194-71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[124] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," *International Journal of Scientific and Research Publications (IJSRP),* vol. 9, no. 10, pp. 143-150, 2019.

[125] I. Kononenko and M. Robnik-Šikonja, "Non-Myopic Feature Quality Evaluation with (R)ReliefF," in *Computational Methods of Feature Selection*, H. Liu and H. Motoda Eds. Boca Raton, Florida: Taylor & Francis Group, 2008.

[126] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, no. 1, pp. 23-69, 2003/10/01 2003, doi: 10.1023/A:1025667309714.

[127] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics,* vol. 85, pp. 189-203, 2018/09/01/ 2018, doi: https://doi.org/10.1016/j.jbi.2018.07.014.

[128] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Aaai*, 1992, vol. 2, no. 1992a, pp. 129-134.

[129] D. Zongker and A. Jain, "Algorithms for feature selection: An evaluation," in *Proceedings of 13th International Conference on Pattern Recognition*, 25-29 Aug. 1996 1996, vol. 2, pp. 18-22 vol.2, doi: 10.1109/ICPR.1996.546716.

[130] M. Tan, J. Pu, and B. Zheng, "Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model," (in eng), *Int J Comput Assist Radiol Surg,* vol. 9, no. 6, pp. 1005-1020, 2014, doi: 10.1007/s11548-014-0992-1.

[131] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification, IEEE Transaction on Systems, Man and Cybernitics, Vol," *SMC,* no. 3 (6), p. 610, 1973.

[132] X. Tang, "Texture information in run-length matrices," (in eng), *IEEE Trans Image Process,* vol. 7, no. 11, pp. 1602-9, 1998, doi: 10.1109/83.725367.

[133] X. Yu, W. Pang, Q. Xu, and M. Liang, "Mammographic image classification with deep fusion learning," *Scientific Reports,* vol. 10, no. 1, p. 14361, 2020/09/01 2020, doi: 10.1038/s41598-020-71431-x.

[134] K. Mendel, H. Li, D. Sheth, and M. Giger, "Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography," (in eng), *Acad Radiol,* vol. 26, no. 6, pp. 735-743, 2019, doi: 10.1016/j.acra.2018.06.019.

[135] M. J. Eppstein and P. Haake, "Very large scale ReliefF for genome-wide association analysis," in *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 15-17 Sept. 2008 2008, pp. 112-119, doi: 10.1109/CIBCB.2008.4675767.

[136] B. Draper, C. Kaito, and J. Bins, "Iterative Relief," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, 16-22 June 2003 2003, vol. 6, pp. 62-62, doi: 10.1109/CVPRW.2003.10065.

[137] J. H. Moore and B. C. White, "Tuning ReliefF for Genome-Wide Genetic Analysis," Berlin, Heidelberg, 2007: Springer Berlin Heidelberg, in Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics, pp. 166-175.

[138] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," (in eng), *IEEE Trans Pattern Anal Mach Intell,* vol. 29, no. 6, pp. 1035-51, Jun 2007, doi: 10.1109/tpami.2007.1093.

[139] W. Yunzhi *et al.*, "A hybrid deep learning approach to predict malignancy of breast lesions using mammograms," in *Proc.SPIE*, 2018, vol. 10579, doi: 10.1117/12.2286555. [Online]. Available: https://doi.org/10.1117/12.2286555

[140] W. Sun, B. Zheng, and W. Qian, "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis," *Computers in Biology and Medicine,* vol. 89, pp. 530-539, 2017/10/01/ 2017, doi: https://doi.org/10.1016/j.compbiomed.2017.04.006.

[141] X. Chen *et al.*, "Development of a transferring GAN based CAD scheme for breast mass classification: an initial study," in *Biophotonics and Immune Responses XVI*, 2021, vol. 11643: SPIE, pp. 15-21.

[142] N. F. Boyd *et al.*, "Mammographic Density and the Risk and Detection of Breast Cancer," *New England Journal of Medicine,* vol. 356, no. 3, pp. 227-236, 2007, doi: 10.1056/NEJMoa062790.

[143] J. Brodersen and V. D. Siersma, "Long-term psychosocial consequences of false-positive screening mammography," *The Annals of Family Medicine,* vol. 11, no. 2, pp. 106-115, 2013.

[144] D. Scutt, G. A. Lancaster, and J. T. Manning, "Breast asymmetry and predisposition to breast cancer," (in eng), *Breast Cancer Res,* vol. 8, no. 2, p. R14, 2006, doi: 10.1186/bcr1388.

[145] M. Tan, W. Qian, J. Pu, H. Liu, and B. Zheng, "A new approach to develop computer-aided detection schemes of digital mammograms," *Physics in Medicine & Biology,* vol. 60, no. 11, p. 4413, 2015.

[146] M. E. Ashgan M. Omer, "Preprocessing of Digital Mammogram Image Based on Otsu’s Threshold," *American Scientific Research Journal for Engineering, Technologym and Sciences,* vol. 37, no. 1, 2017.

[147] J. Chakraborty, S. Mukhopadhyay, V. Singla, N. Khandelwal, and P. Bhattacharyya, "Automatic detection of pectoral muscle using average gradient and shape based feature," (in eng), *J Digit Imaging,* vol. 25, no. 3, pp. 387-99, Jun 2012, doi: 10.1007/s10278-011-9421-y.

[148] M. Jas, S. Mukhopadhyay, J. Chakraborty, A. Sadhu, and N. Khandelwal, "A heuristic approach to automated nipple detection in digital mammograms," (in eng), *J Digit Imaging,* vol. 26, no. 5, pp. 932-40, Oct 2013, doi: 10.1007/s10278-013-9575-x.

[149] B. Zheng *et al.*, "Multiview-based computer-aided detection scheme for breast masses," *Medical Physics,* vol. 33, no. 9, pp. 3135-3143, 2006, doi: https://doi.org/10.1118/1.2237476.

[150] S. Wienbeck *et al.*, "Breast lesion size assessment in mastectomy specimens: Correlation of cone-beam breast-CT, digital breast tomosynthesis and full-field digital mammography with histopathology," (in eng), *Medicine (Baltimore),* vol. 98, no. 37, p. e17082, Sep 2019, doi: 10.1097/md.0000000000017082.

[151] Y. Díez *et al.*, "Revisiting Intensity-Based Image Registration Applied to Mammography," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 5, pp. 716-725, 2011, doi: 10.1109/TITB.2011.2151199.

[152] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," *Frontiers in neuroinformatics,* vol. 7, p. 45, 2013.

[153] C. H. Wei, Y. Li, and C. T. Li, "Effective Extraction of Gabor Features for Adaptive Mammogram Retrieval," in *2007 IEEE International Conference on Multimedia and Expo*, 2-5 July 2007 2007, pp. 1503-1506, doi: 10.1109/ICME.2007.4284947.

[154] M. A. Jones, R. Faiz, Y. Qiu, and B. Zheng, "Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features," (in eng), *Phys Med Biol,* vol. 67, no. 5, Feb 21 2022, doi: 10.1088/1361-6560/ac5297.

[155] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[156] F. Aghaei, M. Tan, A. B. Hollingsworth, and B. Zheng, "Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy," (in eng), *J Magn Reson Imaging,* vol. 44, no. 5, pp. 1099-1106, Nov 2016, doi: 10.1002/jmri.25276.

[157] M. A. Jones, R. Faiz, Y. Qiu, and B. Zheng, "Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features," *Physics in Medicine &amp; Biology,* vol. 67, no. 5, p. 054001, 2022/02/21 2022, doi: 10.1088/1361-6560/ac5297.

[158] R. M. Nishikawa, R. A. Schmidt, M. N. Linver, A. V. Edwards, J. Papaioannou, and M. A. Stull, "Clinically missed cancer: how effectively can radiologists use computer-aided detection?," (in eng), *AJR Am J Roentgenol,* vol. 198, no. 3, pp. 708-16, Mar 2012, doi: 10.2214/ajr.11.6423.

[159] R. Hupse *et al.*, "Computer-aided detection of masses at mammography: interactive decision support versus prompts," (in eng), *Radiology,* vol. 266, no. 1, pp. 123-9, Jan 2013, doi: 10.1148/radiol.12120218.

[160] S. a. A. Hassan, M. S. Sayed, M. I. Abdalla, and M. A. Rashwan, "Breast cancer masses classification using deep convolutional neural networks and transfer learning," *Multimedia Tools and Applications,* vol. 79, no. 41, pp. 30735-30768, 2020.

[161] M. Tan, J. Pu, S. Cheng, H. Liu, and B. Zheng, "Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near-Term Breast Cancer Risk," *Annals of Biomedical Engineering,* vol. 43, no. 10, pp. 2416-2428, 2015/10/01 2015, doi: 10.1007/s10439-015-1316-5.

[162] M. A. Jones, N. Sadeghipour, X. Chen, W. Islam, and B. Zheng, "A multi-stage fusion framework to classify breast lesions using deep learning and radiomics features computed from four-view mammograms," *Medical Physics,* 2023.

[163] S. Li, M. Dong, G. Du, and X. Mu, "Attention dense-u-net for automatic breast mass segmentation in digital mammogram," *IEEE Access,* vol. 7, pp. 59037-59047, 2019.

[164] F. A. Zeiser *et al.*, "Segmentation of masses on mammograms using data augmentation and deep learning," *Journal of digital imaging,* vol. 33, pp. 858-868, 2020.

[165] L. Tsochatzidis, P. Koutla, L. Costaridou, and I. Pratikakis, "Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses," *Computer Methods and Programs in Biomedicine,* vol. 200, p. 105913, 2021.

[166] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015: Springer, pp. 234-241.

[167] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," (in eng), *Biometrics,* vol. 44, no. 3, pp. 837-45, Sep 1988.

[168] N. R. Rajalakshmi, R. Vidhyapriya, N. Elango, and N. Ramesh, "Deeply supervised u-net for mass segmentation in digital mammograms," *International Journal of Imaging Systems and Technology,* vol. 31, no. 1, pp. 59-71, 2021.

[169] T. Shen, C. Gou, J. Wang, and F.-Y. Wang, "Simultaneous segmentation and classification of mass region from mammograms using a mixed-supervision guided deep model," *IEEE Signal Processing Letters,* vol. 27, pp. 196-200, 2019.

[170] M. A. Al-Antari, M. A. Al-Masni, and T.-S. Kim, "Deep learning computer-aided diagnosis for breast lesion in digital mammogram," *Deep Learning in Medical Image Analysis: Challenges and Applications,* pp. 59-72, 2020.

[171] N. Saffari *et al.*, "Fully automated breast density segmentation and classification using deep learning," *Diagnostics,* vol. 10, no. 11, p. 988, 2020.

[172] V. K. Singh *et al.*, "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network," *Expert Systems with Applications,* vol. 139, p. 112855, 2020.

[173] S. Bhattacharjee, S. Poddar, A. Bhaumik, I. K. Maitra, D. Susanna, and A. Ware, "Pre-processing techniques to facilitate better detection

of breast abnormalities using Digital Mammogram," in *AIP Conference Proceedings*, 2023, vol. 2854, no. 1: AIP Publishing.

[174] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," (in eng), *J Digit Imaging,* vol. 26, no. 6, pp. 1045-57, Dec 2013, doi: 10.1007/s10278-013-9622-7.

[175] W. A. Berg *et al.*, "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk," *Jama,* vol. 307, no. 13, pp. 1394-1404, 2012.

[176] B. K. Patel, M. Lobbes, and J. Lewin, "Contrast enhanced spectral mammography: a review," in *Seminars in Ultrasound, CT and MRI*, 2018, vol. 39, no. 1: Elsevier, pp. 70-79.

[177] S. Vedantham, A. Karellas, G. R. Vijayaraghavan, and D. B. Kopans, "Digital breast tomosynthesis: state of the art," *Radiology,* vol. 277, no. 3, p. 663, 2015.

[178] S. T. Taba, T. E. Gureyev, M. Alakhras, S. Lewis, D. Lockie, and P. C. Brennan, "X-ray phase-contrast technology in breast imaging: principles, options, and clinical application," *American Journal of Roentgenology,* vol. 211, no. 1, pp. 133-145, 2018.

[179] N. Berger *et al.*, "Dedicated breast computed tomography with a photon-counting detector: initial results of clinical in vivo imaging," *Investigative radiology,* vol. 54, no. 7, pp. 409-418, 2019.

[180] J. Zuluaga-Gomez, N. Zerhouni, Z. Al Masry, C. Devalland, and C. Varnier, "A survey of breast cancer screening techniques: thermography and electrical impedance tomography," *Journal of medical engineering & technology,* vol. 43, no. 5, pp. 305-322, 2019.

[181] M. F. Covington, E. E. Parent, E. H. Dibble, G. M. Rauch, and A. M. Fowler, "Advances and Future Directions in Molecular Breast Imaging," (in eng), *J Nucl Med,* vol. 63, no. 1, pp. 17-21, Jan 2022, doi: 10.2967/jnumed.121.261988.

[182] N. Utzon-Frank, I. Vejborg, M. von Euler-Chelpin, and E. Lynge, "Balancing sensitivity and specificity: Sixteen year's of experience from the mammography screening programme in Copenhagen, Denmark," *Cancer Epidemiology,* vol. 35, no. 5, pp. 393-398, 2011/10/01/ 2011, doi: https://doi.org/10.1016/j.canep.2010.12.001.

[183] J. J. Fenton *et al.*, "Influence of computer-aided detection on performance of screening mammography," (in eng), *N Engl J Med,* vol. 356, no. 14, pp. 1399-409, Apr 5 2007, doi: 10.1056/NEJMoa066099.

[184] J. G. Elmore and C. I. Lee, "Artificial Intelligence in Medical Imaging—Learning From Past Mistakes in Mammography," *JAMA Health Forum,* vol. 3, no. 2, pp. e215207-e215207, 2022, doi: 10.1001/jamahealthforum.2021.5207.