UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

APPLICATION OF DEEP LEARNING TO OPTIMIZE COMPUTER-AIDED-DETECTION AND DIAGNOSIS OF MEDICAL IMAGES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

Sai Kiran Reddy Maryada

Norman, Oklahoma

2023

APPLICATION OF DEEP LEARNING TO OPTIMIZE

COMPUTER-AIDED-DETECTION AND DIAGNOSIS OF MEDICAL IMAGES

A DISSERTATION APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Dean Hougen, Chair

Dr. Talayeh Razzaghi

Dr. Sridhar Radhakrishnan

Dr. Sivaramakrishnan Lakshmivarahan

Table of Contents

Acknowledgements

Foremost, I extend my deepest appreciation to my advisor, Dr. Dean Hougen, whose unwavering guidance, support, and encouragement have been pivotal throughout my doctoral journey. His profound expertise and insights, coupled with exceptional patience, have been crucial to the sculpting of my research and the accomplishment of my academic objectives. His mentorship extends beyond professional realms, enriching my experience with genuine friendship, for which I am eternally thankful.

My heartfelt thanks are also due to the esteemed members of my current research committee: Dr. Sridhar Radhakrishnan, Dr. Lakshmivarahan Sivaramakrishnan, and Dr. Talayeh Razzaghi. Their distinguished scholarship and supportive presence have not only enriched my academic endeavors but have also provided a framework of mentorship that I deeply honor and value.

A special acknowledgement is owed to Dr. Bin Zheng and Dr. Christan Grant, former members of my committee who have since departed the university. Their early contributions to my research and professional development have left a lasting impact, and I am immensely thankful for the foundation they helped to build in my scholarly pursuits.

I am indebted to my extended family, friends, and colleagues for their enduring patience and understanding amidst times when my focus was divided and my presence scarce. Your unwavering support has been a cornerstone of my resilience and success. The prospect of rekindling our shared moments is one I look forward to with great eagerness.

In closing, an immeasurable amount of gratitude is reserved for my parents, my brother, and my spouse. Your boundless love and belief have been my sanctuary in challenging times.

Abstract

The field of medical imaging informatics has experienced significant advancements with the integration of artificial intelligence (AI), especially in tasks like detecting abnormalities in retinal fundus images. This dissertation focuses on four interrelated research contributions that address crucial aspects of AI in medical imaging, offering a comprehensive overview of various innovative approaches and methodologies. The first contribution involves developing a two-stage deep learning model. This model significantly improves the accuracy of identifying high-quality retinal fundus images by eliminating those with severe artifacts. It highlights the critical role of an optimal training dataset in enhancing the performance of deep learning models. The second contribution presents an innovative algorithm for synthetic data generation. This algorithm enhances the effectiveness of deep learning models in medical image analysis by augmenting datasets with synthesized annotated diseased regions onto disease-free images, leading to notable improvements in disease classification accuracy. The third contribution is centered around a novel joint deep-learning model for medical image segmentation and classification. Combining a U-net architecture with an image classification model it demonstrates substantial accuracy improvements as the training dataset size increases. Lastly, a comparative analysis is conducted between radionics-based and deep transfer learning-based Computer-Aided Detection (CAD) schemes for classifying breast lesions in digital mammograms. The findings reveal the superiority of deep transfer learning methods in achieving higher classification accuracy. Collectively, these contributions offer valuable insights and practical methodologies for enhancing the efficiency and diagnostic accuracy of AI applications in medical imaging, marking a significant step forward in this rapidly evolving field.

1. Introduction to Deep Learning to Optimize CAD of Medical Images.

Background

In modern medicine, the convergence of cutting-edge technology and clinical practice has given rise to transformative tools that can revolutionize diagnostic precision and enhance patient care. Computer-aided diagnosis (CAD) systems are central to this paradigm shift, which harnesses the power of sophisticated algorithms to assist healthcare practitioners in interpreting medical images.

Computer-Aided Diagnosis (CAD): These systems serve as an invaluable adjunct to human expertise, offering a meticulous analysis of complex medical imagery [1]–[3]. By scrutinizing the minutiae within images, CAD systems bring to light subtle anomalies that might otherwise escape human observation. In this regard, CAD has emerged as a vital ally, empowering diagnosticians with a second set of expert eyes, particularly in fields where accuracy is paramount.

Deep Learning in Medical Imaging Informatics: At the vanguard of CAD lies the integration of deep learning models, a subset of artificial intelligence that excels at tasks requiring pattern recognition within vast and complex datasets [4]–[6]. Convolutional Neural Networks (CNNs) are at the heart of these models, adept at autonomously extracting intricate features from raw data. This hierarchical learning capability enables deep learning models to discern nuanced relationships that may elude human perception. These models have demonstrated unparalleled success across many applications, including natural language processing, signal processing, and medical image analysis.

Convolutional Neural Networks (CNNs): CNNs are deep learning models designed explicitly for processing grid-structured data, such as images [7]–[9]. They have revolutionized the field of computer vision and are widely used in tasks like image recognition, object detection, and

segmentation. They are inspired by the organization of the visual cortex in the human brain and are designed to automatically and adaptively learn hierarchical representations of data through the use of convolutional layers.

Here are some key terminologies associated with CNNs:

1. Convolutional Layers:

- CNNs derive their name from the operation they apply, known as convolution. This operation involves sliding a filter (also known as a kernel) over the input image. The filter is a small matrix of weights convolved with a region of the input image. This process allows the network to learn features like edges, corners, textures, and more complex patterns at different levels of abstraction.

2. Filters and Feature Maps:

- Filters in a CNN play a critical role in learning relevant features from the input data. Each filter detects a particular feature type, like edges, textures, or complex patterns. As the network trains, these filters evolve to recognize increasingly sophisticated features.
- The output of applying a filter to a region of the input is called a feature map. Multiple filters are applied to the input image in parallel to generate various feature maps. These maps collectively represent the activation of different features across the image.

3. Activation Function:

- After convolution, each element in the feature map goes through an activation function, usually a Rectified Linear Unit (ReLU). ReLU introduces non-linearity, allowing the network to learn complex relationships between features.

4. Pooling Layers:

- Pooling layers help down-sampling the feature maps' spatial dimensions while retaining important information. The most common pooling operation is max pooling, which selects the maximum value from a local region. This reduces the spatial size, decreases the computational load, and helps make the model more robust to variations in object position.

5. Fully Connected Layers:

- After several convolutional and pooling layers, the data is typically flattened and fed into fully connected layers. These layers connect every neuron from one layer to every neuron in the next layer, enabling the network to learn complex relationships between features.

6. Dropout:

- Dropout is a regularization technique commonly used in CNNs. It involves randomly deactivating a fraction of neurons during training. This prevents overfitting by ensuring that no single neuron becomes overly dependent on specific features.

7. Backpropagation:

- CNNs are usually trained using backpropagation, an algorithm that adjusts the weights of the filters and fully connected layers based on the error (difference between predicted and actual output) calculated during each training iteration.

CNNs excel at learning hierarchical features. Lower layers learn simple features like edges, while deeper layers learn more complex features like textures and object parts. This hierarchical representation allows CNNs to recognize intricate patterns in images.

The architecture and parameters of a CNN, including the number of layers, filter sizes, and strides, are carefully designed to suit the specific task. By combining these elements, CNNs have proven to be remarkably effective in a wide range of computer vision applications.

The Pioneering Role of Deep Learning in Medical Imaging: The studies outlined in this dissertation embody the pioneering spirit that deep learning has instilled in medical imaging informatics. Each study stands as a testament to the potential of artificial intelligence to unravel complexities within medical imagery and illuminate novel paths to more accurate diagnoses.

Retinal fundus photography, initially a focal point of ophthalmology, has garnered widespread interest for its diagnostic potential beyond ocular health. It offers a unique window into the broader realm of neurologic disorders and systemic diseases, revealing vital information about a patient's overall health. Yet, the quest for large, clean, diverse datasets remains a formidable challenge. The research presented in the first study addresses this pressing concern by introducing a sophisticated two-stage deep learning model. This model identifies pristine retinal fundus images and effectively filters out those marred by artifacts, substantially improving classification accuracy [10].

The second study, encapsulated in the paper "A Novel and Efficient Synthetic Data Generation Algorithm to Improve Efficacy of Deep Learning Models of Medical Images" [11], addresses the perennial challenge of dataset acquisition. Through ingenuity, this study pioneers an algorithm capable of synthesizing images, thus augmenting datasets with invaluable diversity and circumventing the limitations posed by privacy regulations and the labor-intensive process of manual curation. The algorithm synthesizes diseased regions onto disease-free images, significantly expanding the dataset and resulting in a notable boost in disease classification.

The third study, detailed in the paper “Improving Medical Image Segmentation and Classification Using a Novel Joint Deep Learning Model” [12], introduces a joint model that marries U-net architecture with image classification. This fusion of capabilities transcends conventional segmentation and classification tasks, promising heightened accuracy, particularly as the size of the training dataset expands.

The fourth study, articulated in the paper “A Comparison of Computer-Aided Diagnosis Schemes Optimized Using Radiomics and Deep Transfer Learning Methods” [13], presents a comprehensive comparative analysis between the radiomics-based and deep transfer learning-based CAD schemes. Through rigorous investigation, we demonstrate that deep transfer learning offers a more efficient avenue to develop CAD schemes, yielding higher classification accuracy than its radiomics-based counterpart.

Collectively, these studies witness a future where CAD systems, fortified by the prowess of deep learning models, are poised to usher in a new era of diagnostic precision and efficiency. They not only represent significant contributions to the scientific discourse but also hold the promise of practical integration into clinical practice, ultimately enhancing patient outcomes.

2. Applying a Novel Two-Stage Deep-Learning Model to Improve Accuracy in Detecting Retinal Fundus Images

2.1 Motivation for Two-Stage Deep-Learning Model

In medical image analysis, clinical data representation and extraction are the primary purposes and the premise of many complicated frameworks. Many models have been developed in the field [14]–[16]. These models assist clinicians or researchers in simplifying and evaluating medical images in various applications, e.g., organ segmentation [17]–[19], anatomical landmark detection [20], [21], and computer-aided diagnosis (CAD) [5], [22], [23]. Various organ systems have highly diverse attributes and medical image models are usually designed for specific studies to assimilate appropriate prior knowledge. However, models using handcrafted features or parameter settings are often difficult to optimize and not easily adaptable to other CAD schemes and applications.

Thus, in the last decade, deep-learning technology and models have emerged and attracted extensive research and development interest in broad scientific communities. Deep-learning models [24] have performed successfully in various applications, including natural-language processing [25], signal processing [26], and object detection [9]. Many variants of deep learning architectures addressing diabetic retinopathy were also proposed [27]. Their popularity comes from their ability to model nonlinear and hierarchical features from many complex datasets. Deep network architectures are able to extract features under supervised or unsupervised environment from large quantities of training data. Image recognition is a promising application where deep learning learns good visual features. Likewise, deep learning may be put to good use for medical image recognition as well.

A retinal fundus photograph consists of a color image of an eye's interior surface. Such photographs help ophthalmologists identify disorders. Examining retinal fundus (ocular fundus) imagery has become increasingly valuable to other medical fields besides ophthalmology. Neurologic disorder early indications can be found using optic nerve, retina, and choroid abnormalities [28]. Such imagery also helps neurologists understand underlying systemic diseases that contribute to a patient's neurological condition. Hence, fundus photography can help physicians to identify severe conditions. Developing computed aided diagnosis (CAD) models using artificial intelligence (AI) or deep machine learning (ML) with superior performance to identify such underlying conditions requires a large, clean, and diverse dataset. Due to HIPAA and other regulations, these datasets can be difficult to acquire from private sources.

Using public data platforms such as the Google search engine is one method to obtain such a dataset. However, when building the dataset using appropriate keywords in the search engine, many non-useful and noisy images with severe artifacts or overlapped un-related items are also returned along with the useful fundus photography images. To select useful and clean fundus images visually or manually is a tedious and time-consuming task.

In order to improve efficacy of training and applying deep learning models using relatively small image datasets, the objective of this study is to investigate the feasibility and advantages of developing a new automated scheme to identify and select useful retinal fundus images from all downloaded images without human intervention. Each downloaded image is to be categorized into one of three classes based on its content: Class 1 – complete, single retinal fundus images without annotations; Class 2 – single images containing multiple fundus photos or single fundus photos with annotations; Class 3 – noisy images overlapped with artifacts. Class 1 is the target class that

goes into the database. In this study, it is essential to achieve the highest possible PPV for Class 1 images.

2.2 Materials and Methods

2.2.1 Image Dataset from Google Search Engine

We use the Google search engine to query and retrieve rental fundus photos or images. A total of 1,227 unique images are downloaded from the Google search engine. These images are downloaded in batches using different keywords related to retinal fundus images. Each image is classified into one of three classes. Each image in Class 1 is a single fundus photo without any annotations. Each image in Class 2 is a single fundus photo with annotations or multiple fundus photos with at least 25% of the total area of the image containing a fundus photo, as shown in Figure 2-2. Class 3 consists of images in which the fundus photos comprise less than 25% total area of the image and other noisy images, as shown in Figure 2-3. Based on above criteria, Class 1 has 620 images, Class 2 has 134 images, and Class 3 has 473 images.

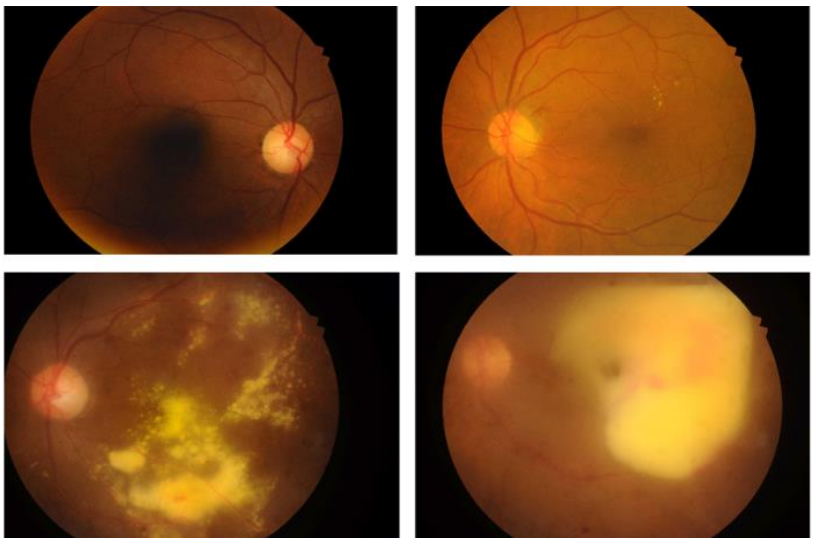

Figure 2-1: Four example images of Class 1.

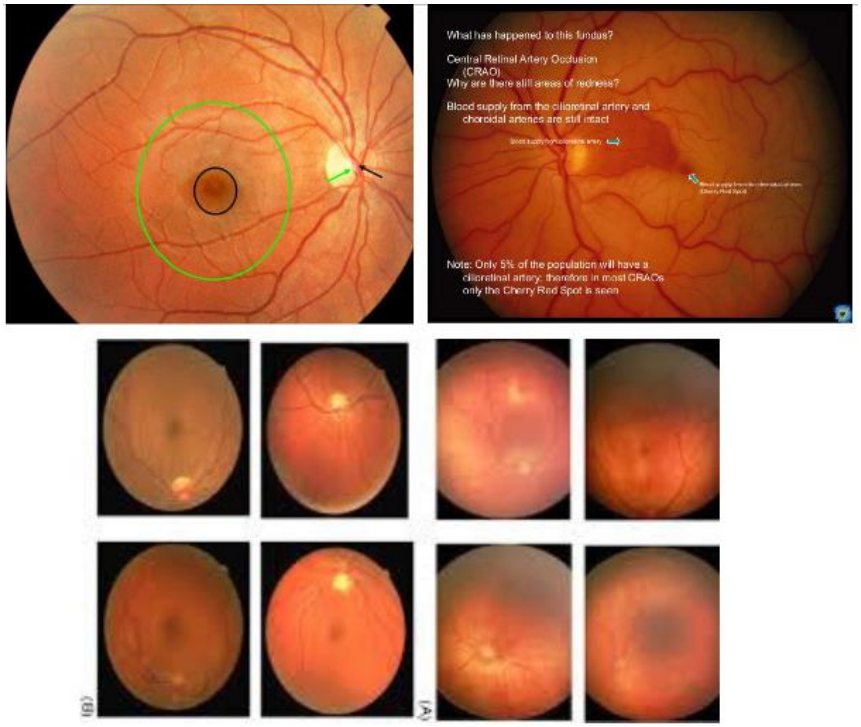

Figure 2-2: A sample of Class 2 images. Two on the top and one on the bottom.

Figure 2-3: A sample of Class 3 images.

We use transfer learning to build new deep-learning models to classify the retinal fundus images into 3 classes. For this purpose, the ResNet-50 [7] architecture is selected to build our transfer learning model, which has been pre-trained using a large ImageNet database. Although many Deep Convolutional Neural Network (DCNN) models have been developed and used as transfer learning models in the medical imaging informatics field, we chose a DCNN model using ResNet-50 architecture. ResNet-50 has several advantages compared with other architectures like VGG19 [29] and AlexNet [28] such as a smaller number of parameters to train (more depth, less width), reduced effect of vanishing gradient, and higher accuracy obtained on the ImageNet dataset.

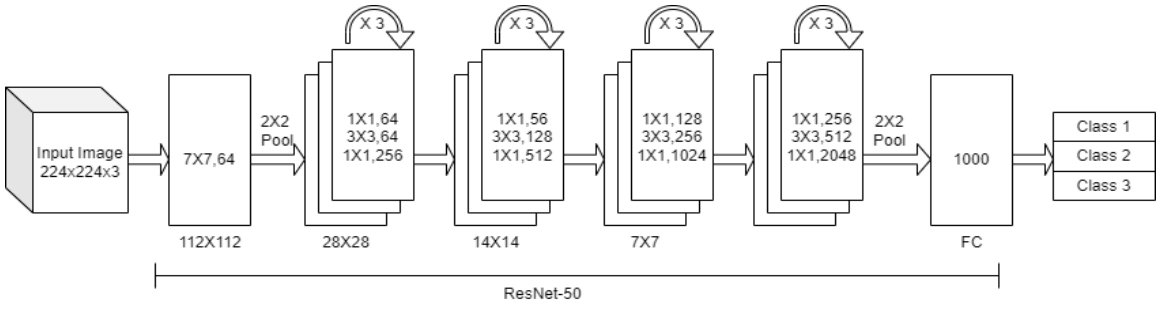

Figure 2-4: Block diagram of stage 1 transfer learning with ResNet-50.

In our study, we first randomly select 50 images from each of the three classes to build a training dataset. Thus, a total of 150 images are used to fine-tune the ResNet-50 model. The fine-tuned ResNet-50 model is then applied to the remaining 1077 images to classify these images into three classes. Since the model generates three probability scores in three classes, the image is assigned to the class that has the highest probability score.

However, we recognize that the conventional classification method shown in Figure 2-4 based solely on the highest probability score for three classes may not achieve optimal classification

performance. Thus, in order to optimize model performance, we propose a new method of classification shown in Figure 2-2 by adding a threshold value on the SoftMax activation function in the last layer of the trained ResNet-50 model. The idea behind increasing the threshold (0.5 to 0.9) is to identify the optimal threshold value that can yield the highest positive predictive value (PPV) in classification and minimize false positives. Specifically, after adding a threshold, a test image will only be assigned to a class in which the model-generated probability score is greater than the threshold.

After adding a threshold to SoftMax, a group of images will become undetermined if all 3 probability scores are smaller than the threshold. These undetermined images are assigned to a new class namely, Class 4, representing the "difficult" images. In order to further classify these difficult images, we add a second transfer learning model. The same pre-trained ResNet-50 model is fine-tuned again using part of these difficult images. As a result, a unique two-stage model is built. The model in the first stage is applied to identify and classify "easy" images. If the test images are classified as undetermined difficult images, they will be further analyzed and classified by another model in the second stage.

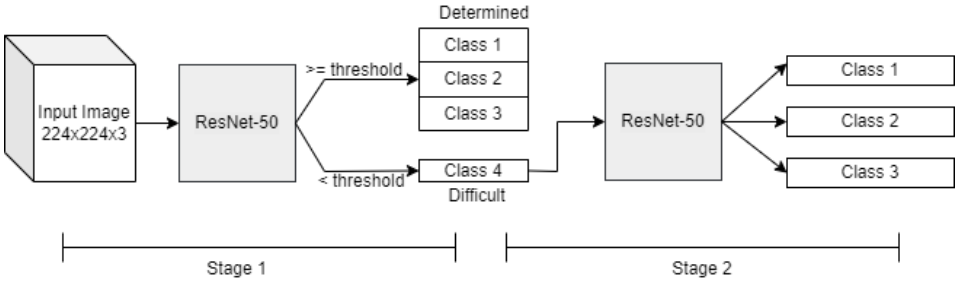

Figure 2-5: Block diagram of the proposed 2-stage model.

Due to the limitations of the dataset size used to test the above two-stage concept, we use 5-fold cross-validation to train (fine-tune) and test the ResNet-50 model in the second stage. In both stages, the pre-trained ResNet-50 architecture is used for training. All the images are normalized

and resized to 224×224×3 to fit into the pre-trained ResNet-50 network. Horizontal and vertical flipping is used for augmentation.

2.3 Performance of Two-Stage CNN vs Single-Stage CNN:

Table 2-1 shows performance when using one ResNet-50 model to classify images into 3 classes

Class	1	2	3	Total
Assigned images	588	109	380	1077
True positive (TP) images	563	60	367	990
False positive (FP) images	25	49	13	87
Positive predictive value (PPV)	0.957	0.550	0.966	0.919

without adding a threshold value on the SoftMax layer. Over all 1077 test images, the model predicts and assigns 588, 109 and 380 images to Classes 1, 2, and 3, respectively. Among the classified images, true- and false-positives along with the positive predictive values are summarized in

Table 2-1.

Class	1	2	3	Total
Assigned images	588	109	380	1077
True positive (TP) images	563	60	367	990
False positive (FP) images	25	49	13	87
Positive predictive value (PPV)	0.957	0.550	0.966	0.919

Positive predictive value (PPV)	0.957	0.550	0.966	0.919

Table 2-1: Stage 1 results without threshold on the SoftMax layer

After adding threshold values on SoftMax layer of the model, the easy images will be classified into three classes and the difficult images will be assigned into Class 4 (undetermined images). As the threshold values increase from 0.5 to 0.9, the results show that PPV values of Class 1 increase, while the number of determined images decreases due to the increase of undetermined images in Class 1 (as shown in *Figure 2-4* and *Figure 2-5*, and *Table 2-2*).

Table 2-2: Performance of model in Stage 1 using different threshold values on the SoftMax layer.

Threshold value	0.5	0.6	0.7	0.8	0.9
Number of determined images	1034	939	799	561	234
True positive (TP) images	965	887	768	544	226
False positive (FP) images	69	52	31	17	8
PPV of 3 Classes	0.933	0.945	0.961	0.970	0.966

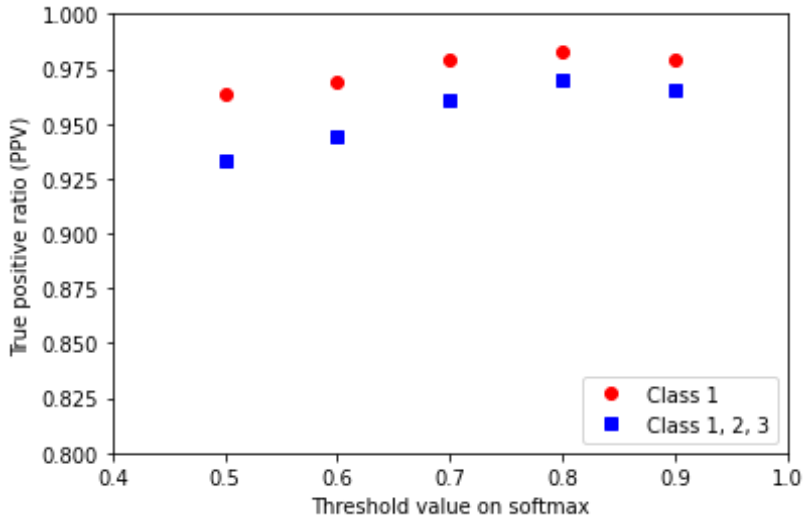

Figure 2-6: True positive ratio (PPV) VS Threshold value.

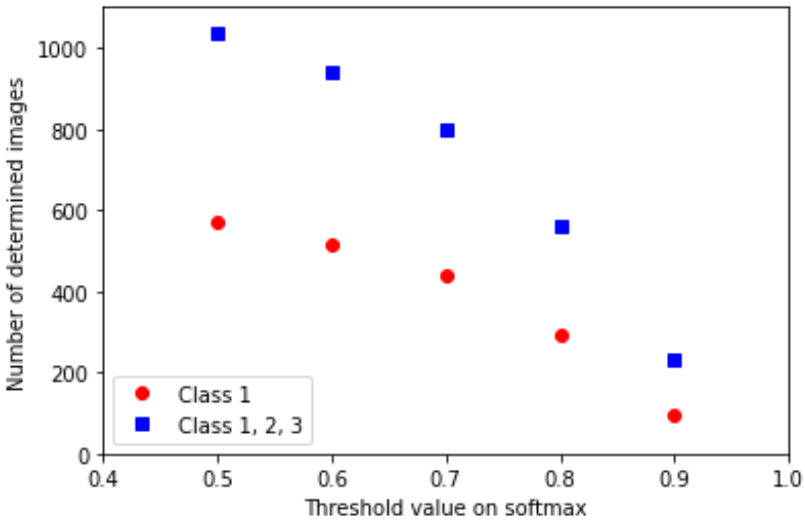

Figure 2-7: Number of determined images VS Threshold value.

Table 2-3 summarizes the performance of the ResNet-50 model implemented in Stage 2 to classify the difficult images, which are undetermined by the Stage 1 model based on the SoftMax score

with a threshold value of 0.8. The undetermined/difficult images (1077-561 = 516) are used in Stage 2.

Table 2-3: Performance of model in Stage 2 with different threshold values on the SoftMax layer

Threshold value	0.5	0.6	0.7	0.8	0.9
Number of determined images	516	513	508	504	486
True positive (TP) images	490	477	489	485	476
False positive (FP) images	25	36	19	19	10
PPV of 3 Classes	0.950	0.930	0.963	0.962	0.979

Table 2-4 shows the classification results using the two-stage model with threshold value = 0.8 on the SoftMax layer of the ResNet-50 model in the first stage. For example, the model in Stage 1 assigns 293 images to Class 1 and the model in Stage 2 assigns 278 images to Class 1. After running this two-stage model, a total of 571 images are assigned to Class 1. The positive predictive value (PPV) is 0.986. In comparing single-stage with multi-stage models (Table 2-4 and Table 2-1), we observe that false positives are reduced from 25 to 8 for Class 1 (the target class representing complete retinal fundus image without annotations), and from 87 to 36 for all three classes. The PPV of Class 1 increases from 95.7% to 98.6%, and the PPV over all three classes increases from 91.9% to 96.6% using two-stage model. The number of true positives for all three classes increases from 990 to 1,029 using the two-stage model. However, the total number of true positives for Class 1 remains the same using the single-stage and two-stage models.

Table 2-4: Classification performance using the new two-stage model with a threshold of 0.8 on the SoftMax layer of the model in the first stage.

Class	1	2	3	Total
Assigned images	571	77	417	1065
True positive (TP) images	563	52	410	1029
False positive (FP) images	8	25	**7**	36
Positive predictive value	0.986	0.675	0.9832	0.966

2.4 Discussion:

In this study, we proposed, developed, and tested a new two-stage model in classifying retinal fundus images using transfer learning ResNet-50 models, based on our experimental results, we have the following key observations.

- First, using a two-stage transfer learning model outperforms a traditional single stage model. For instance, while observing the key statistical parameters (
- Table 2-1 and Table 2-4) for the entire dataset, we noticed an improvement in the PPV

Class	1	2	3	Total
Assigned images	588	109	380	1077
True positive (TP) images	563	60	367	990
False positive (FP) images	25	49	13	87
Positive predictive value (PPV)	0.957	0.550	0.966	0.919

value from 91.9% to 96.6%. This is an increase of 4.7% as compared to the base model.

- Second, our study primarily focuses on identifying complete retinal fundus images (Class-1). On comparing the results of Class-1 between a single stage and two-stage transfer learning models, even though the total number of images identified is equal, a significant reduction in the false positives is noticed (from 25 to 8).
- Third, the first stage of our proposed model is essential in identifying the optimal threshold to improve the potential samples required for Stage 2. From Figure 2-6 and Table 2-2, we can see that increasing the threshold from 0.8 to 0.9 doesn't contribute to increasing the PPV value.
- Additionally, we also observed that like conventional machine learning or computer-aided detection schemes, performance of a deep-learning model also heavily depends on the content distribution of training datasets. Thus, it is often difficult to correctly identify or classify the difficult or subtle images using a deep-learning model that is trained or fine-tuned using a small dataset which does not contain or cannot sufficiently represent the difficult images (or outliers).
- In order to optimally address this challenge, developing a two or multi-stage deep-learning model or scheme has significant advantages. This is the primary contribution of this study to the medical imaging informatics or CAD field.

Last, we also recognize that many previous studies have highlighted the significance of integrating the image processing techniques along with the deep learning frameworks. Thus, in our future studies, we plan to explore this phenomenon using a more diverse dataset of retinal images and investigate the application of the integrated architectures that combine both traditional image processing and deep learning frameworks.

3 An Efficient Synthetic Data Generation Algorithm to Improve Efficacy of Deep Learning Models of Medical Images

3.1 Motivation for Synthetic Data Generation Algorithm

Medical images are widely used to detect and diagnose diseases. However, reading and interpreting medical images by clinicians is difficult and time-consuming, which results in large intra- and inter-observer variability [10], [30]–[33]. Thus, computer-aided detection and diagnosis (CAD) schemes have been developed aiming to improve the diagnostic accuracy of medical images. CAD schemes are used to detect diseases, classify disease types or severities, and predict disease prognosis or response to treatment [10], [32]–[38]. Due to errors in lesion segmentation and identification of effective features, CAD performance and clinical utility are limited.

To overcome these limitations, new CAD models using deep convolutional neural networks (DCNNs) have been widely investigated and used in CAD schemes of medical images [39]–[41]. The advantages of using DCNN models over traditional machine learning classifiers based on handcrafted features to develop CAD systems for medical images have been demonstrated empirically [41]. Deep-learning models in CAD schemes for medical images require large, diverse datasets to reduce bias and increase robustness. For example, a study using deep learning to detect diabetic retinopathy using retinal fundus photographs reported a high detection accuracy of 0.991 AUC [42].

However, developing new CAD schemes using DCNN models also faces a challenge due to the limited size of image datasets and lack of resources to retrieve and label images. The existing datasets are often imbalanced with fewer positive cases. It is also impractical to ask clinicians to review and annotate large numbers of images for research purposes [43]. Thus, two approaches of

using data augmentation and transfer learning are commonly applied to develop DCNN models in the medical imaging field. Several studies [44] have shown that fine-tuning DCCN models, which were originally trained on a large and diverse dataset of natural images (i.e., ImageNet [8]), using small datasets of medical images helped improve accuracy and decreased time for convergence of model training or fine-tuning.

Despite potential advantages, data augmentation and deep transfer learning can only partially reduce the impact of small image datasets in medical imaging, as they cannot increase the diversity of medical image cases. Thus, to find a better solution, researchers have also investigated whether adding synthesized data [45] generated from real medical image data can help overcome the challenges and develop more robust DCNN models. One recent study [46] shows that the synthetic images generated by a generative adversarial network (GAN) were not recognized as synthetic by radiologists and using these GAN-generated synthetic images increased performance of a DCNN model in classifying liver lesions. Thus, using GANs is shown to be a promising solution to generate more synthetic image data. Thus, other studies have reported that using GANs to generate synthetic images of skin lesions [47], colon mucosa [48], and chest X-ray images [49].

In our studies, we recognized that GANs are quite complex and not well-investigated for generating diverse medical images to date. To address this issue, we propose to investigate and test a new algorithm that generates synthetic medical images without GANs by extracting annotated diseased regions, redistributing them, and projecting them onto negative (healthy) images. Our hypothesis is that this method can include more diverse negative images that do not need expert annotations, resulting in synthetic image data with a broad diversity of normal tissue patterns, potentially helping to better train DCNN models and improve performance. Therefore, the objective of this study is to validate the feasibility of our hypothesis by comparing 3 deep

transfer learning models (VGG-16 [29], ResNet-50 [7] and Inception V3 [50]) trained using a small set of original images and a large set of the synthetic images, respectively. Details of this study are reported in the following sections.

3.2 Materials and Methods

3.2.1 IDRiD Image Dataset

In this study, we first used the Indian Diabetic Retinopathy Image Dataset (IDRiD), a publicly available dataset of retinal fundus images [27], which has been used in a previous international competition of using deep learning models to segment lesions. The dataset includes retinal images acquired from 81 patients along with corresponding annotated binary masks for three types of lesions: Hemorrhages (HE), Hard Exudates (EX), and Soft Exudates (SE). In this dataset, all three types of lesions were annotated in the images by clinicians. This dataset is randomly divided into two independent subsets, with 54 retinal fundus images selected for training and 27 images for testing.

In this dataset, each patient falls into one of two categories. In *Category 1*, an image contains all three types of lesions, while in *Category 2*, an image contains only HE and EX lesions. In the training subset, 26 and 28 images belong to Categories 1 and 2, respectively. In the testing subset, 14 and 13 images are assigned to Categories 1 and 2, respectively. Each case includes one original retinal fundus image and three or two annotated mask images, one for each type of disease in Categories 1 and 2, respectively. Figure 3-1 shows an example image in Category 1 and three corresponding masks of HE, EX, and SE lesions. Figure 3-2 shows another example image belonging to Category 2 along with the two associated masks of HE and EX lesions. All labeled mask images are pseudo binary images without lesion density information.

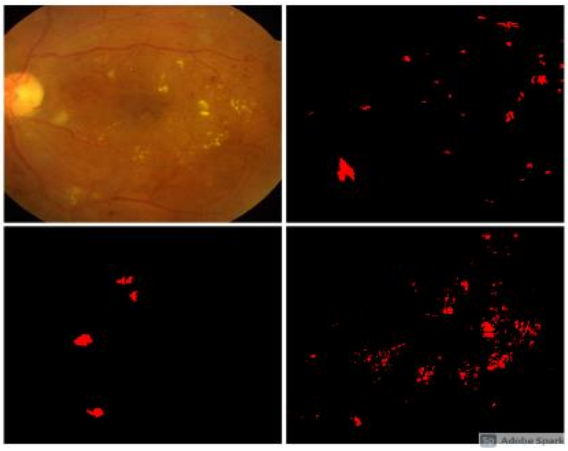

Figure 3-1: Original retinal image (top left) with three annotated masks for three types of lesions (top right-clockwise: HE, EX, and SE).

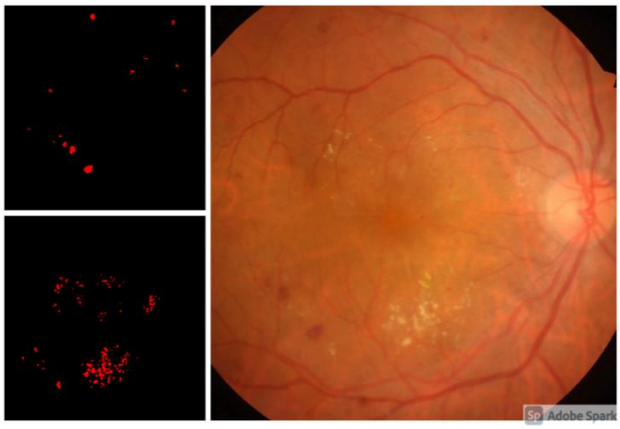

Figure 3-2: Original retinal image (right side) with two annotated masks for two types of lesions (top left: HE; bottom left: EX).

To prepare images for use in the study, we applied the following image preprocessing steps. First, edge detection was performed horizontally on each retinal fundus image to find the maximum width of the fundus within the image. Second, the image was cropped to this width to remove excess background from the image. Third, a black (background) padding method was applied vertically to convert the image into a square. Last, each image was resized to a square region of

interest (ROI) of 225×225 pixels, which matches the input image size of the VGG-16 and ResNet-50 DCNN models. Figure 3-3 shows an example of this image preprocessing. The dark background regions on the left and right side are removed, and two padding strips are added on the top and bottom of the images to convert the image into a square ROI with 225×225 pixels. Similarly, we have added extra padding to generate another set of images with square ROI of size 299×299 pixels, which is the input image size for Inception V3.

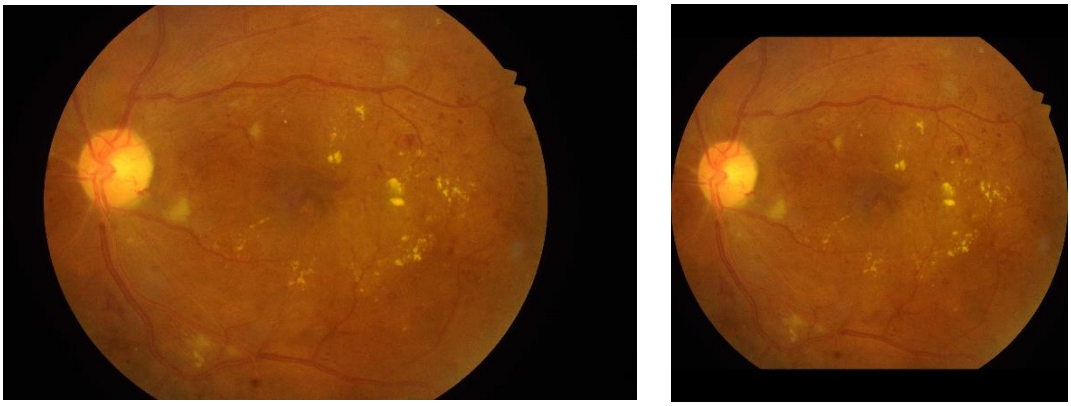

Figure 3-3: Original retinal image (left side); image after pre-processing (right side).

3.2.2 Synthetic Data Generation Algorithm

To generate synthetic images for this study, we acquired another dataset of 60 healthy (or normal) retinal fundus images, provided by an ophthalmologist as a base to generate images with synthetically added diseased lesions. Then we designed and applied the following synthetic data generation algorithm to extract lesion blobs, randomize their distribution to create new lesions, and project the synthetic lesions onto these healthy retinal fundus images.

1. Because there are three separate disease masks of HE, EX, and SE lesions in Category 1 images (as shown in Figure 3-1) and two separate disease masks of HE and EX lesions in Category 2

images (as shown in Figure 3-2) in the IDRiD dataset, we created a combined mask image that contains either three types of lesions for the Category 1 disease case or both HE and EX lesions for the Category 2 disease case in the training subset by using element-wise addition, which means inserting the lesions extracted in the original masks into a uniform "black" mask. The top row of Figure 3-4 shows one combined mask that contains all three types of lesions extracted from one Category 1 disease case. As a result, this step creates 54 combined masks, in which 26 represent masks of Category 1 cases and 28 represent Category 2 cases.

2. After creating the combined mask image of each case, we count the number of lesion blobs (N) in each mask image and label them from 1 to N. After examination of whole dataset, we noticed that the counted number of lesion blobs (N) varies from a minimum of 20 to more than 100 in different mask images. Depending on the number of lesion blobs (N) associated with each case, we randomly select 2 to N blobs from each originally combined mask to generate 20 new masks. Thus, in this step, we generate 1,080 new mask images (54 combined masks × 20 new masks per combined mask) that each contain a random number of lesion blobs. Figure 3-4 shows an example of one originally combined lesion mask and 20 new lesion masks generated from this combined lesion mask.

3. We use element-wise multiplication to identify the lesion type of each blob in the mask randomly generated in Step 2. The results are saved in a database that records the location of each blob and the associated lesion type (HE, EX, or SE). Since the masks generated in Step 2 are pseudo-binary mask images, in this step, we map them into new mask images that contain real lesion blobs. For this purpose, we project the pseudo blobs in the masks generated in Step 2 back to the original retinal fundus images to extract real lesion blobs.

4. Each real lesion blob mask obtained in Step 3 is flipped both horizontally and vertically. One real lesion mask becomes three masks (original, horizontally flipped, and vertically flipped) with the same set of real lesion blobs located in different positions and orientations. Thus, we generate a total of 3,240 real lesion masks (1,080 generated mask images × 3 transformation options). Due to random selection process, the final real lesion blob masks can be divided into seven categories containing lesions of (1) HE+SE+EX, (2) HE+SE, (3) EX+SE, (4) HE+EX, (5) HE, (6) SE, (7) EX. Among these masks, most of them contain EX+SE and HE+SE+EX categories; the number of masks in the other five categories are much smaller. In order to overcome this data imbalance issue, we performed multiple angular rotations to increase the numbers of masks in the five categories to generate approximately comparable numbers. Table 3-1 shows the resulting number of the final synthetic lesion blob masks in each of the seven categories.

5. Finally, we insert and project each of the lesion blob masks generated in Step 4 onto a randomly chosen healthy retinal fundus image. In the projection, the overlapping pixels between the healthy retinal fundus image and the synthetic lesion blob mask are taken from the synthetic lesion blob mask, whereas the remaining pixels are extracted from the healthy fundus image. As a result, we generated 7,092 synthetic images of Category 1 and 6,786 synthetic images of Category 2, as shown in Table 1, which were then used to train deep learning model as described below.

Table 3-1: Numbers of synthetic real lesion blob masks generated in seven distributions.

Category	Lesion composition	Number of images
1	HE, SE, and EX	7092

2	HE and SE	6786
3	EX and SE	6939
4	HE and EX	7902
5	HE	7641
6	SE	6780
7	EX	6780

In summary, Figure 3-5 illustrates the step-by-step workflow of the proposed synthetic image data generation algorithm in which the figure shows an original retinal fundus image (Figure 3-5 5a) and corresponding lesion mask (Figure 5b) these are taken from IDRiD dataset, a randomly generated lesion mask (Figure 5c), a lesion blob image that is obtained after projecting the mask with the corresponding original image (Figure 5d), a sample flip (horizontal, Figure 5e), a random angular orientation (Figure 5f), and a lesion blob mask insertion to a healthy retinal fundus image to generate the final synthetic image (Figure 5g). Figure 3-6 demonstrates nine synthetic images on the right side and one magnified view of one synthetic image with three clusters of blobs annotated on the left side.

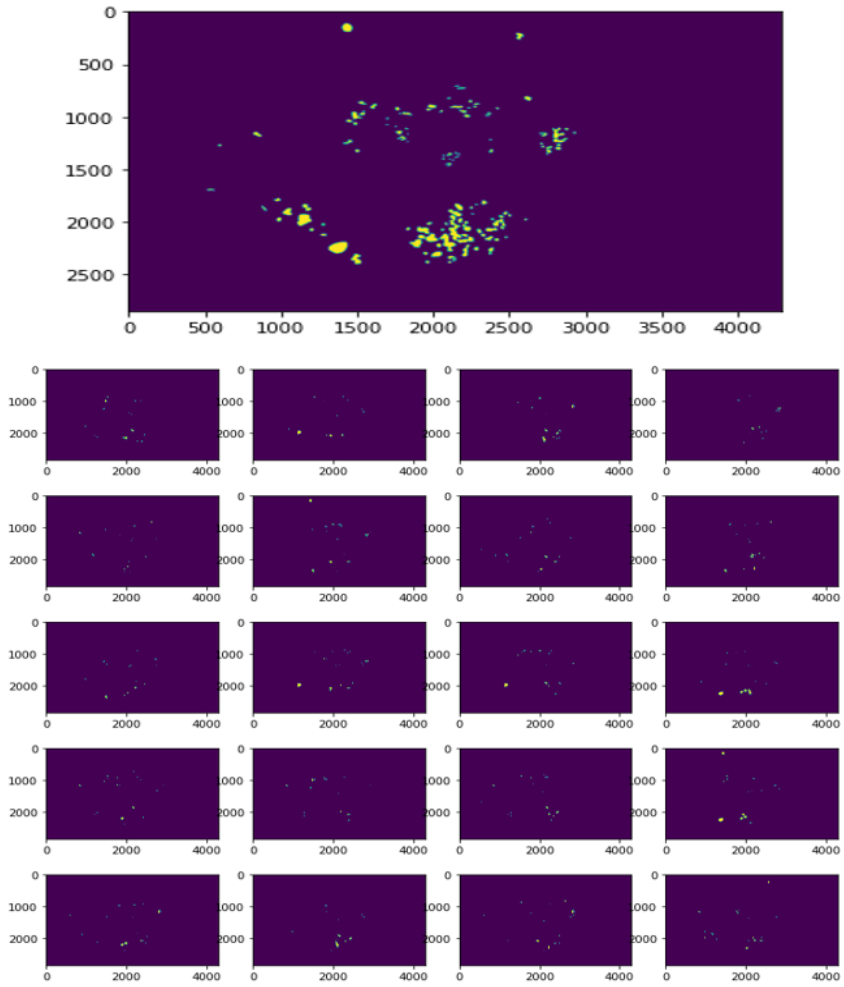

Figure 3-4: The image on top shows one originally combined mask image that consists of all 3 types of lesions (Category 1) and images below on the grid show 20 new mask images generated based on the top combined mask image with random distributions of lesion blobs.

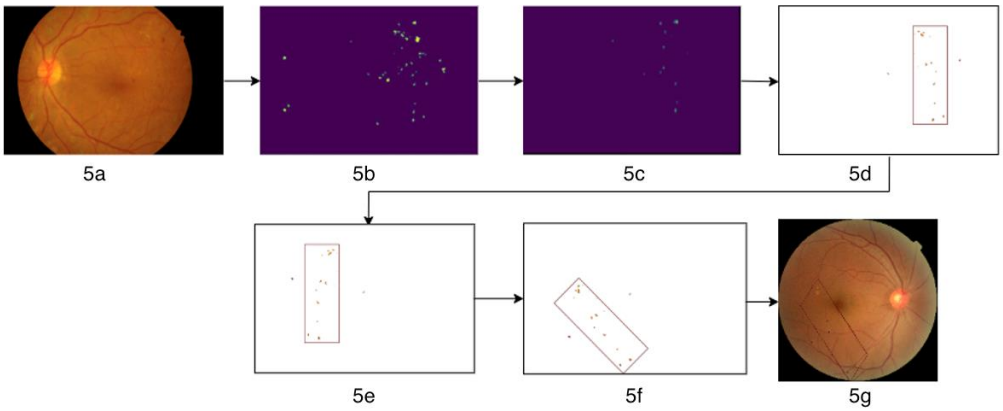

Figure 3-5: A detailed step-by-step illustration of the proposed algorithm.

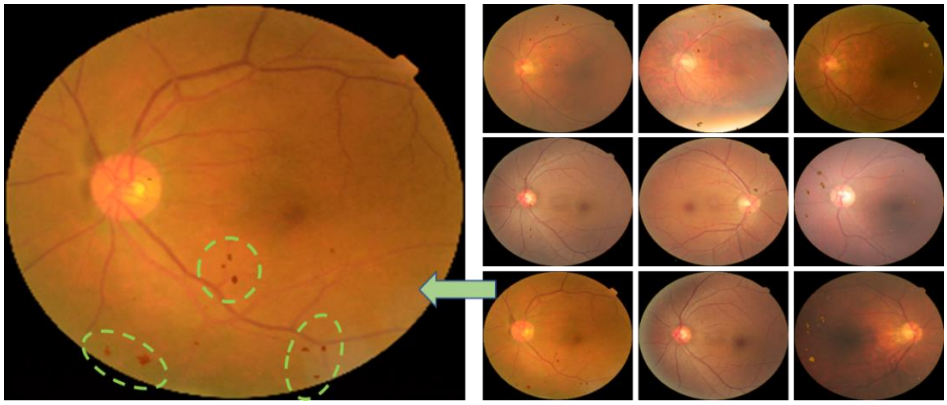

Figure 3-6: Synthetic data after projecting lesions onto healthy images (right side); magnified view of one sample case with some lesions annotated (left side).

3.2.3 Experiments

In this study, we chose three popular DCNN models namely, VGG16, ResNet-50, and Inception-v3 as base models to build six transfer learning DCNN models to classify retinal fundus images into two disease categories. Using each base model, two transfer learning DCNN models were

trained using the original image data provided in the IDRiD dataset and the synthetic image data generated by the new algorithm developed in this study, respectively. Model Set 1 and Model Set 2 where each model set contains the three base models, Model-1 was trained using 26 Category 1 images and 28 Category 2 images acquired from 54 original IDRiD images, while Model-2 was trained using 7,092 Category 1 synthetic images and 6,786 Category 2 synthetic images. Figure 3-7 show the deep learning architecture of three modified VGG-16, ResNet-50 and Inception-v3 models, respectively. All retinal fundus images were resized to fit the input image size of each model (224×224 pixels for VGG-16 and ResNet-50 models and 299×299 for Inception-3 model).

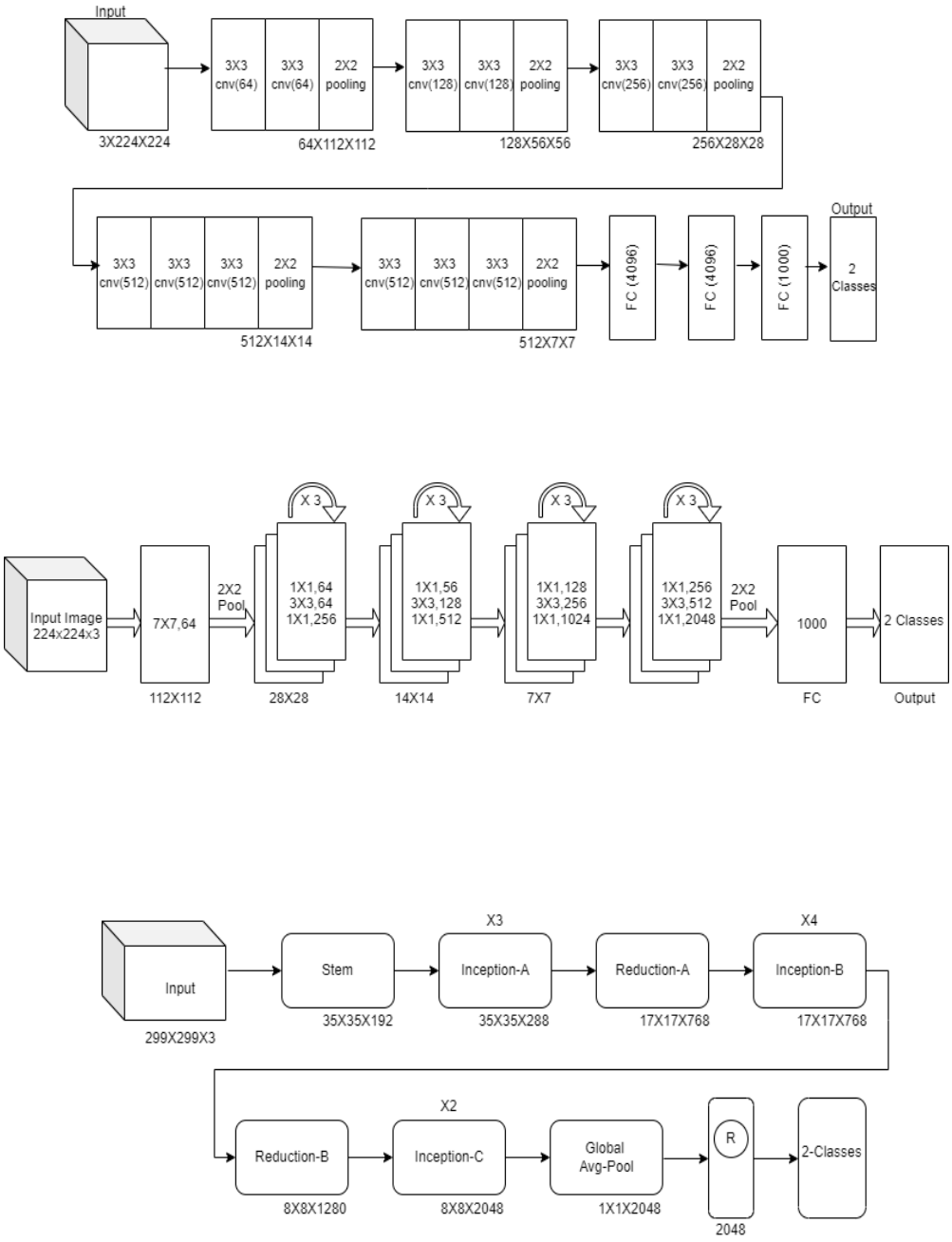

Figure 3-7: Illustration of the modified three transfer learning DCNN models of VGG-16 (top), ResNet-50 (middle) and Inception-v3 network (bottom).

Using either real or synthetic retinal fundus images, the VGG-16, ResNet-500, and Inception-v3 based DCNN models were fine-tuned to classify images as Category 1 or Category 2 retinal

diseases. In the fine-tuning process, the Adam optimizer is adapted to use a variable learning rate that starts from 0.006 and exponentially decays by a factor of 0.05 for every three epochs. To minimize or reduce overfitting risk, models are trained using 25 epochs based on the cross-entropy loss of the Adam optimizer. Then, the last SoftMax layer of each model is modified to one output neuron with sigmoid activation that achieve the classifying two categories of retinal fundus images in this study.

After fine-tuning, each model was applied to the same independent test dataset of 27 images. Since the last SoftMax layer has two output nodes that generate two probability scores indicating the likelihood of a test image belonging to two disease categories, the test image is assigned to the category that has the higher probability score. From test results, we calculated true positive (TP), false negative (FN), true negative (TN), and false positive (FP) values to generate a confusion matrix, from which we evaluate model classification performance by computing four commonly used evaluation indices namely, accuracy, precision, recall, and F1 score, using Equations (3.1) – (3.4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.4}$$

Last, classification performance indices of the different DCNN models separately trained using the original IDRiD images and the synthetic images were tabulated and compared.

3.3 Model Performances Using Original Dataset and Synthetic Dataset

Figure 3-8 shows two diagrams that plot curves of disease classification accuracy over 25 training epochs of the VGG-16 transfer learning model trained using the original IDRiD images and the synthetic images, respectively. Similar training epoch patterns are also observed in training ResNet-50 and Inception-3 models. Specifically, in training or fine-tuning each model, the training dataset was divided into two subsets in which 80% of the image data was used to train the model and 20% of image data was used to validate the performance of the model. The results of two trend curves generated by the training and validation data indicate that the classification accuracy approaches a plateau and gradually saturates once epochs exceed 20. Thus, in this study, we chose 25 training epochs for all models.

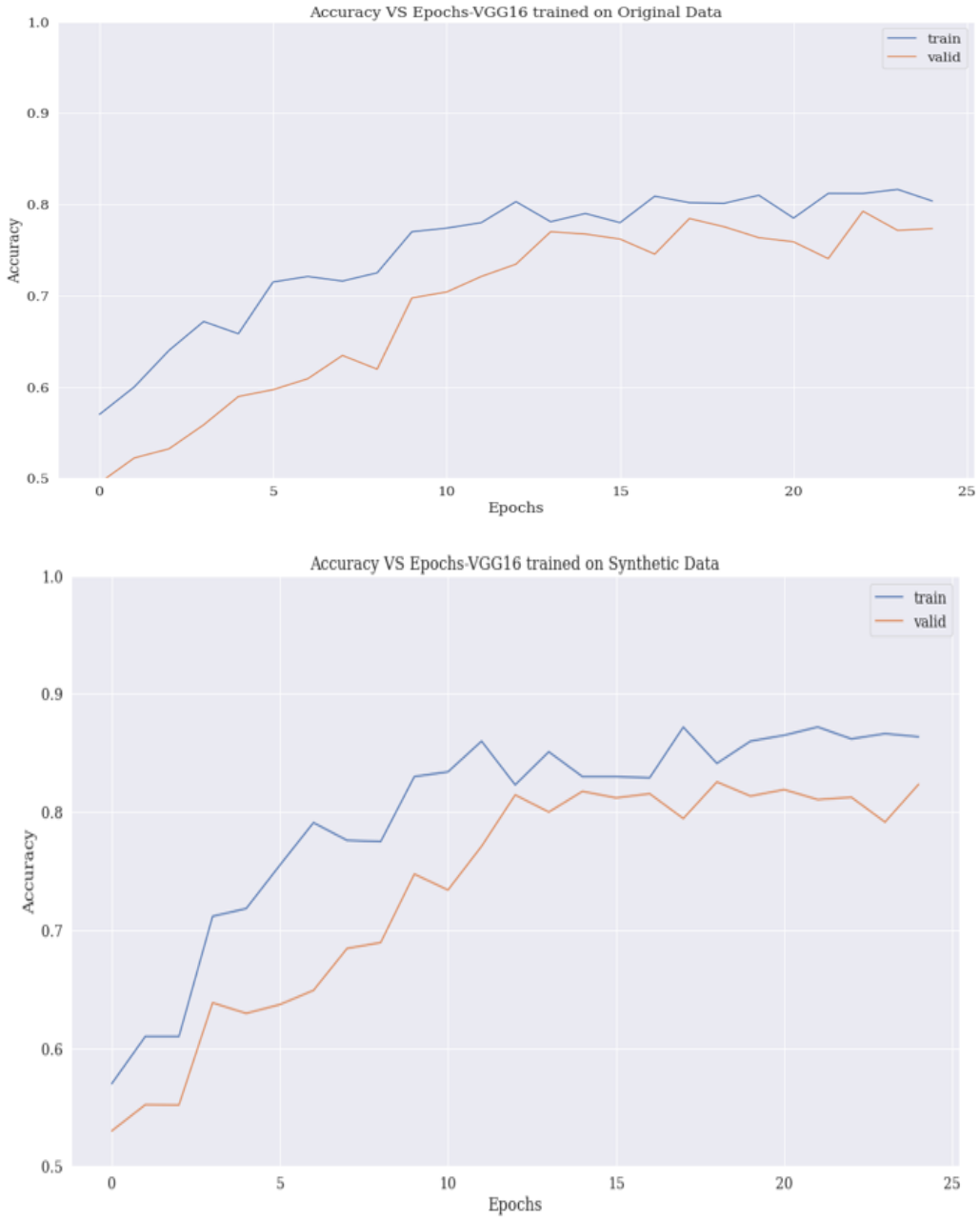

Figure 3-8: Illustration of training and validation accuracy curves of one VGG-16 deep transfer learning model trained using original IDRiD images (top) and synthetic images (bottom).

The testing dataset includes 14 images in Category 1, which includes Hemorrhages, Hard Exudates, and Soft Exudates, and 13 images in Category 2, which includes Hemorrhages and Hard Exudates. Figure 3-9 demonstrates three sets of two confusion matrices generated by applying

three transfer learning DCNN models (VGG-16, ResNet-50, and Inception-v3) trained using the original IDRiD images and the synthetic images to the same testing dataset of 27 images to classify between Category 1 and Category 2 diseases, respectively. Based on these confusion matrices, Table 3-2 shows and compares four evaluation indices of classification performance generated by three deep transfer learning models trained using the original IDRiD images and the synthetic images. The results show that all three transfer learning models (VGG-16, ResNet-50, and Inception-v3) yield an overall classification accuracy of 81.5% (22/27), which is 7.4% higher than the classification accuracy of 74.1% (20/27) using all the three models (VGG-16, ResNet-50, and Inception-v3) trained on the original IDRiD images.

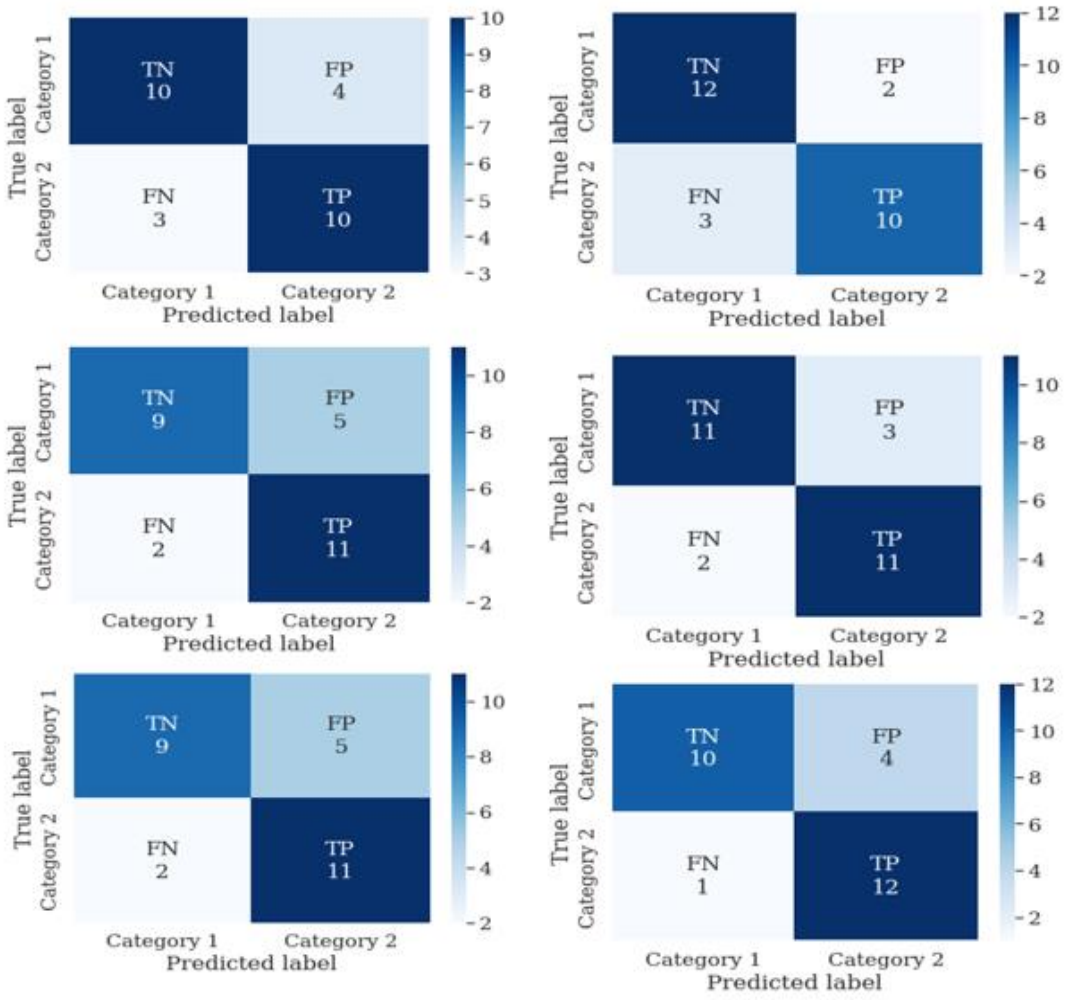

Figure 3-9: Six confusion matrices generated by VGG16 (top), ResNet-50 (middle) and Inception-v3 (bottom) trained using original IDRiD images (left) and synthetic images (right).

Table 3-2: Comparison of various performance metrics while using original and synthetic data.

Model	Training Images	Accuracy	Precision	Recall	F1 Score
VGG-16	Original	0.741	0.714	0.769	0.741
	Synthetic	**0.815**	**0.833**	0.769	0.800
ResNet-50	Original	0.741	0.688	0.846	0.759
	Synthetic	**0.815**	0.786	0.846	0.815
Inception-v3	Original	0.741	0.688	0.846	0.759
	Synthetic	**0.815**	0.750	**0.923**	**0.828**

Although all three models (VGG-16, ResNet-50, and Inception-v3) yield the same classification accuracy, the distribution of other three evaluation indices (precision, recall, and F1 score) are different among the three models. The transfer learning models fine-tuned using the original IDRiD images yield substantially higher precision the models fine-tuned using the synthetic images. The results indicate that the models fine-tuned using synthetic images achieve higher sensitivity and generally higher false positive rates. Combining precision and recall indices, the models fine-tuned using synthetic images yield substantially higher F1 scores in this study. For example, by comparing two Inception-v3 transfer learning models, testing results indicate that using synthetic data to fine-turn the model increases F1 score by 6.9% (from 0.759 to 0.828).

3.4 Discussion

In this study, we demonstrate a new algorithm to generate synthetic retinal fundus images embedded with different types of diseased lesions. This study has several unique characteristics

and interesting observations. First, unlike other existing synthetic image data generation methods (i.e., using a Monte Carlo simulation or a Generative Adversarial Network), which are quite difficult to design and computationally expensive because large numbers of algorithm or network parameters need to be chosen and optimized, our new method is much simpler and more computationally efficient. It uses a multi-stage approach to directly extract positive lesions, randomize distribution of lesion blobs, and then inserts the positive lesion blobs in the randomly selected locations of negative images. In the medical imaging field, collecting large numbers of negative images is much easier than collecting positive images with manual annotation of diseased lesions. In this study, we significantly expanded our dataset size from 54 positive and 60 negative images to 14,994 synthetic images (for Categories 1 and 2 diseases, as shown in Table 1).

Second, although the original 54 images are divided into only two categories, using this new algorithm can generate 49,920 synthetic images in 7 categories of lesions, as shown in Table 3-1. Thus, despite the small image base of 54 positive images and a large number of synthetic images, all algorithm-generated synthetic images have different positive lesion blob combinations, and the lesion blobs randomly distribute in different locations with different orientations, which increases the diversity of training samples by minimizing redundancy of the generated synthetic images.

Third, we fine-tuned three sets of DCNN models based on VGG16, ResNet-50 and Inception-V3 model. Each DCNN set includes two models trained using original IDRiD images (Model-1) and synthetic images (Model-2). Applying to the same testing images, Model-2 yields 7.4% higher classification accuracy than Model-1. Although this accuracy improvement is observed on a quite small testing dataset of 27 images and further test is needed using much large image datasets in the future, the results demonstrate feasibility and advantages of using this new algorithm to generate synthetic images to help improve model performance.

Fourth, classification performance as measured by four evaluation indices (as shown in Table 3-2) is independent from three deep learning DCNN models (VGG-16, ResNet-50 and Inception-V3). It is promising that all three types of DCNN models that are fine-tuned using synthetic images achieve substantially higher F1 scores ranging from 5.6% (ResNet-50) to 6.9% (Inception-v3). Such results indicate the higher robustness of applying this new simple multi-step algorithm to generate synthetic images, which can effectively help train or fine-tune different DCNN models.

Finally, we recognize that seamless insertion of lesions or other abnormality regions onto negative images using this simple algorithm has restrictions including that (1) the lesions must have clear boundary contours so that the lesions can be easily extracted, (2) the normal tissue background should also be relatively uniform. Retinal fundus images meet these two restrictions. Some other medical images (i.e., liver tumors) can also meet these restrictions. Nonetheless, if lesions have fuzzy boundary embedded under heterogenous tissue background (i.e., breast tumors depict on mammograms), this algorithm will not work "as is" and modifications will be needed. Despite such limitations, developing this new simple algorithm to generate synthetic images has its higher clinical impact at least for retinal fundus images that are acquired using a low-cost image examination method and widely used in clinics to screen, detect, and diagnose many common human diseases including eye diseases and diabetes. We will further test and validate this new algorithm and apply it to develop more accurate and robust DCNN models for different medical applications in the future.

4 Improving Medical Image Segmentation and Classification Using a Novel Joint Deep Learning Model

4.1 Motivation for Building Joint Deep Learning Model

Medical imaging analysis deals with the extraction and analysis of clinical data. Many models are developed to achieve this objective [14]–[16], [20], [21]. The primary purpose of these models is to assist clinicians and researchers in a wide range of applications, from organ segmentation [17]–[19] to disease diagnosis. Most models are developed using machine learning during the initial research stages and require domain expertise for feature selection to build better systems [5], [51]–[53]. Also, models developed using the handcrafted features are difficult to adapt to build other computer-aided-diagnosis (CAD) models.

Deep learning successfully navigated the shortcomings mentioned above and attracted extensive research in the past decade. Deep learning has enabled advancements in many fields [54]–[57], including computer vision. Computer vision is a collaborative scientific field that deals with how computers can be trained to gain a high-level understanding of digital images or videos. In recent years, deep learning models have been developed on a wide range of medical image types to diagnose many diseases. In addition, deep learning models are combined and trained to achieve higher efficiency in image processing [58]–[62]. These hybrid or joint models can process different data types and perform better than individually trained models [63]–[67]. Collaborative models learn different versions of features from the data and boost performance. Thus, the insights processed and extracted from joint models aid in solving complex problems. Especially in the field

of biomedical imaging, analyzing features of the data could help medical professionals to create a clear intuition of the issue at hand.

On the other hand, limited access to data in medical domains hinders the application of artificial intelligence. Due to data regulations and market fragmentation, it can be challenging to large, homogeneous datasets. Therefore, creating deep learning models that can work with datasets of limited size is still challenging. Alternatively, gathering additional information from a single image is another approach. For instance, in a traditional binary classification task, each image is designated a single 0/1 label. In supervised tasks, this label and input image are the only information given to a model to optimize its internal parameters. Thus, in small datasets, a lack of knowledge leads to overfitting.

Providing additional information to the model by producing a segmentation mask of the input image can alleviate the overfitting problem. This mask typically has the same dimensionality as the input image and provides significantly more comprehension for the model to optimize. The supplementary information from the mask can help boost model performance and prevent overfitting. Hence, it is possible to improve model performance even with smaller datasets by combining classification and segmentation tasks as a single joint network. Y-Net [69] is one such model that takes full advantage of this approach to joint training. However, the network has not gained widespread appeal, partially because its effectiveness has yet to be demonstrated in independent, controlled studies. This dissertation presents a hybrid deep-learning J-net model that simultaneously generates segmentation masks and diagnoses classification labels for skin cancer images. Our network provides a simple architecture that can be easily adapted to any dataset. Due to the straightforward nature of our network, smaller datasets with a few hundreds of data records can yield highly accurate results that avoids overfitting.

4.2 Materials and Methods

4.2.1 Skin Cancer Image Dataset

This experiment is tested on a dataset obtained from Kaggle, an online open-source platform for data. The dataset has skin cancer images with corresponding segmentation masks and a binary diagnostic label for the pigmented lesions on the skin. The segmentation masks are gray-scale images with white pixels (255) identified as the lesions on the skin and black pixels (0) as the normal skin. The binary label states the presence of melanocytic nevi (NV) in the skin images. This data field has 0 representing ‘mild’ and 1 representing the ‘severe’ presence of nevus cells in the corresponding image of skin cancer lesion. In total, 1200 images are selected with equal distribution of both cases for the study.

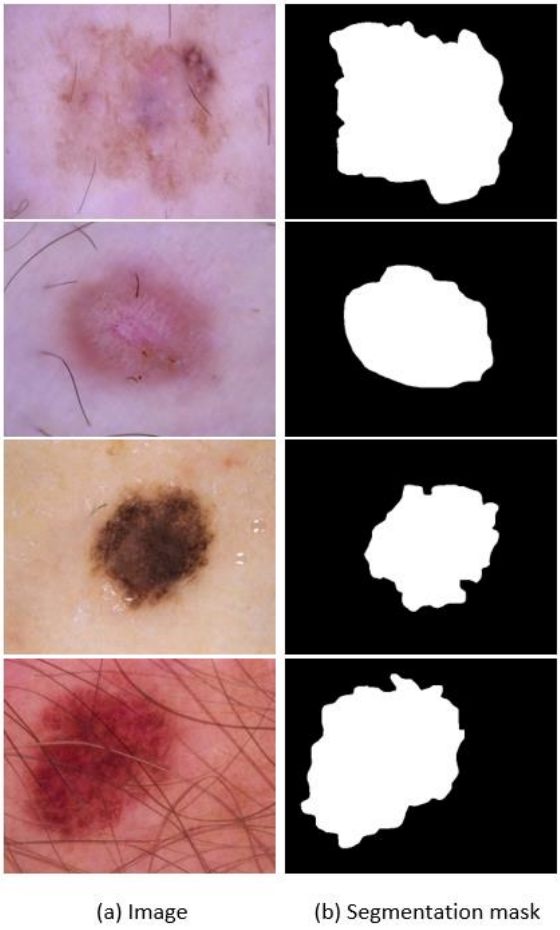

Figure 4-1: Skin cancer images and corresponding lesion masks.

4.2.2 Comparison Between Existing U-Net and Proposed J-Net Architectures:

We compare our novel J-Net architecture to a traditional U-Net architecture (described in Subsubsection 4.2.2.1). For the training of both U-Net and J-Net models, the dataset is split into train, validation, and test ratio of 70:20:10. The images and masks are normalized for data standardization and converted to the size of 256×256. Both models are trained for 150 epochs with a learning rate of 3e-5 and batch size 10 with an SGD optimizer. For the segmentation task, a custom loss function is developed using a sigmoid function followed by a binary cross entropy (BCE) loss function. To optimize the binary classification task, a BCE loss function is used.

4.2.2.1 U-Net

The U-Net architecture was developed for biomedical image segmentation that contains two phases [68]. The first phase is the contraction phase, also known as the *encoder*, which captures the context of the image. The encoder is a traditional stack of convolutional and max-pooling layers. The second phase is the symmetric expanding phase, also known as the *decoder*, which enables precise localization using transposed convolutions. Thus, it is an end-to-end, fully convolutional network (FCN).

For this study, the traditional U-Net architecture is modified by reducing the convolution block size to 32×32 on all the stages of the architecture. This modification reduced the complexity of the model and thus helps to prevent overfitting. Each stage of the encoder has two convolution blocks, where each block has a 3×3 convolution layer followed by a batch normalization and a ReLU activation layer. A 2×2 max pooling layer is used to reduce the dimensionality of the features and passed to the next stage of the encoder. Four encoder stages reduce the image features from 256×256 to 16×16, followed by a bottleneck layer where the features are upsampled and send into the decoder. The decoder blocks upscale the features from the bottom stage and pair with the corresponding encoder features. Following the decoder, a sigmoid layer is used to produce a binary segmentation mask.

How the Decoder Works:

1. **Upsampling**: The decoder in U-Net performs *upsampling* of the feature maps received from the encoder. This is essentially the inverse of the *downsampling* (convolution and pooling) operations that occur in the encoder.
2. **Concatenation with Feature Maps from Encoder**: After each upsampling step, the decoder concatenates the upsampled feature maps with the corresponding feature maps from the encoder. This process is known as *skip connections* and helps the network to retain fine-grained details that might be lost during downsampling.
3. **Convolution Operations**: Following the concatenation, standard convolution operations are applied. These convolutions refine the upsampled features and help in learning localization information necessary for precise segmentation.

Mathematics of Upsampling (Inverse Convolution):

1. **Convolution Transpose:** The mathematical operation used for upsampling is often called "transposed convolution" or "deconvolution". However, it should be noted that this is not a true mathematical inverse of the convolution but rather a way to increase the spatial dimensions of the input feature maps.
2. **Kernel and Stride:** Just like in convolution, transposed convolution also uses a kernel (or filter). The stride in transposed convolution defines how much the output size is increased. A stride of 2, for example, would roughly double the dimensions of the feature map.
3. **Padding and Output Size:** The padding used can affect the output size. Careful calculation of the padding ensures that the output size matches the expected dimensions after upsampling.

4. **Mathematical Operation:** The transposed convolution involves placing the kernel at each element of the input, multiplying, and summing the overlapped values. This creates an output matrix larger than the input.
5. **Backpropagation:** During training, the gradients are computed through these layers in a way that mirrors the forward pass, helping the network to learn the upsampling weights effectively.

How interactions happen between input image and filter in upsampling?

1. Start with the Input Image and Filter:

Input Image:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Upsampling Filter:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

2. First Interaction:
    - Multiply the top-left pixel of the image (value = 1) by the entire 2×2 filter. This will create a 2×2 matrix where all values are 1 (since 1×1=1).
    - Place this 2×2 matrix in the top-left corner of the output image.
3. Move to the Next Pixel:
    - Slide the filter to the right (one pixel over on the input image). Now it covers the second pixel in the top row (value = 2).

- Repeat the multiplication with the filter. This creates another 2×2 matrix with all values as 2.
- Overlay this 2×2 matrix onto the output image, starting from the second row and column. Since there's overlap with the previous step, add the overlapping values.

4. Continue Across the Image:
    - Continue this process across the image, moving the filter right and then down, and overlaying the resulting 2×2 matrices onto the output image.
5. Final Output:
    - Once you've covered all pixels of the input image, you'll have the final upsampled image.

Upsampled Image:

$$\begin{bmatrix} 1 & 3 & 5 & 3 & 0 \\ 5 & 12 & 16 & 9 & 0 \\ 11 & 24 & 28 & 15 & 0 \\ 7 & 15 & 17 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In summary, the decoder in U-Net uses a series of transposed convolutions (upsampling) and regular convolutions, combined with skip connections from the encoder, to reconstruct the image for segmentation tasks. The mathematics of upsampling are similar to convolution but are designed to increase the spatial dimensions of the feature maps.

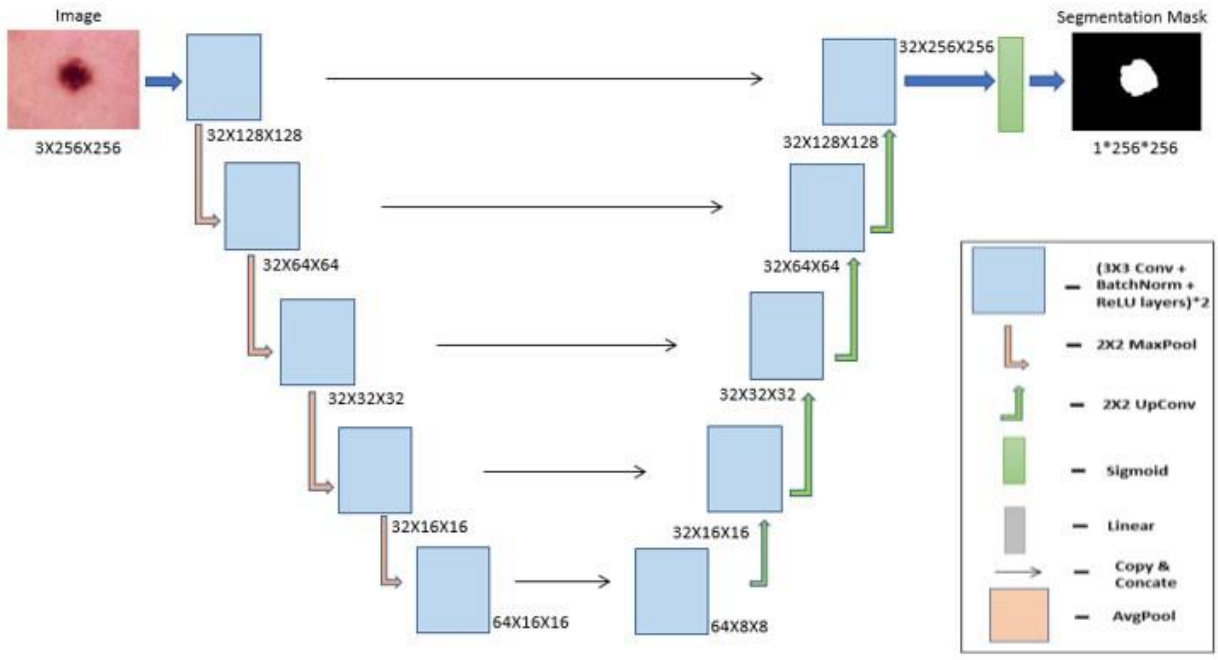

Figure 4-2: U-Net architecture.

4.2.2.2 Binary Image Classifier

The second independent model used to perform image classification is a simple binary classifier. The classifier's architecture is developed to resemble the encoder part of the U-Net plus an additional convolution block followed by an average pool layer and linear layers producing a binary output. The output '1' indicates severe and '0' indicates mild presence of melanoma on the lesions. For the model training, a dataset of 1200 images with a 70:20:10 data spit for training, validation, and test sets is drawn and the performance metrics are measured to compare with the results from the J-Net.

4.2.2.3 J-Net

A novel model, J-Net, is developed by combining the U-Net along with the binary image classifier, inspired by Y-Net [69]. The J-Net architecture contains two-way feature learning modules. The

first module is a U-Net, a deep downsampling-to-upsampling sub-network for semantic features. The second is a convolutional sub-network without downsampling for classification features. The first module has the exact architecture of the U-net described in the previous section. The classification branch attached to the encoder part of U-Net uses the same convolution blocks as the encoder stages of the segmentation branch. An average pooling, three linear, and sigmoid layers are connected to the classification convolution blocks. Thus, the feature maps from the encoder phase of J-Net serves as the input to the binary classifier the classification branch categorizes the image into two classes.

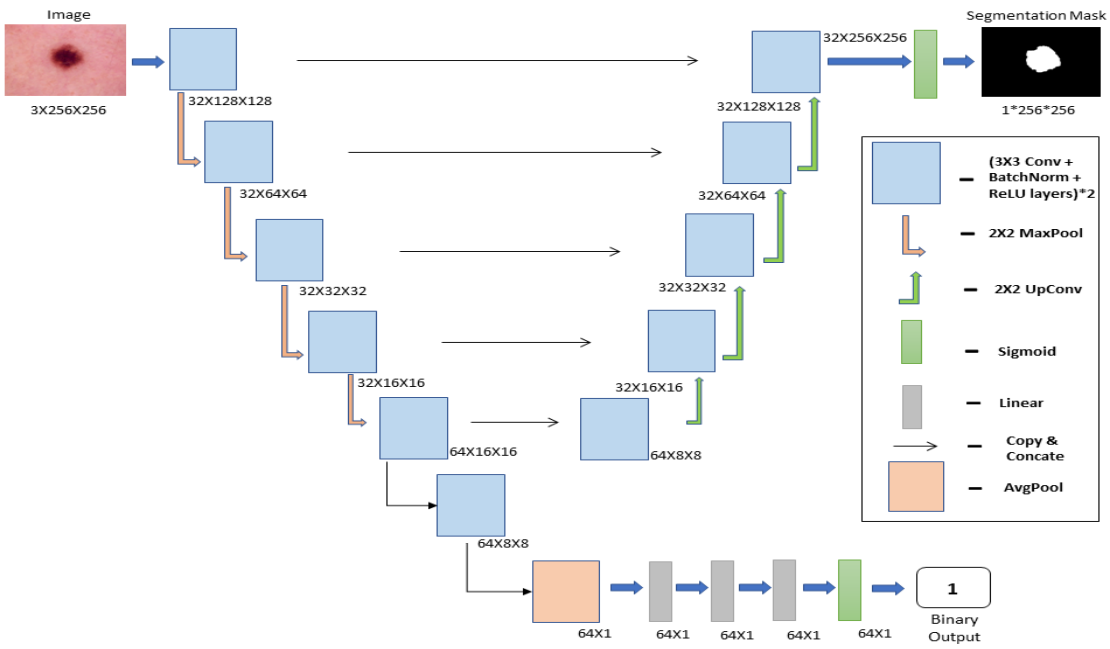

Figure 4-3: J-Net architecture.

The J-Net is optimized using the combined loss from segmentation ($Loss_{Seg}$) and classification tasks ($Loss_{Clas}$). Combining losses helps the network to adapt the features of each data label along with its segmentation mask. The additional information learned by the classification branch helps the segmentation branch in optimizing the masks and vice versa.

$$Loss_{JNet} = Loss_{Seg} + Loss_{Clas} \quad (1)$$

4.3 U-Net vs J-Net model performance comparisons:

The results of the U-Net and the J-Net models are compared based on lesion segmentation and classification performance metrics. For testing purpose, 10 percent of the dataset is used. From the observed results, J-Net has higher segmentation accuracy than U-Net over smaller datasets. Especially, J-Net accurately segments the contours of the images. Both models' segmentation metrics improved as the dataset size increased from 200 to 1200 images.

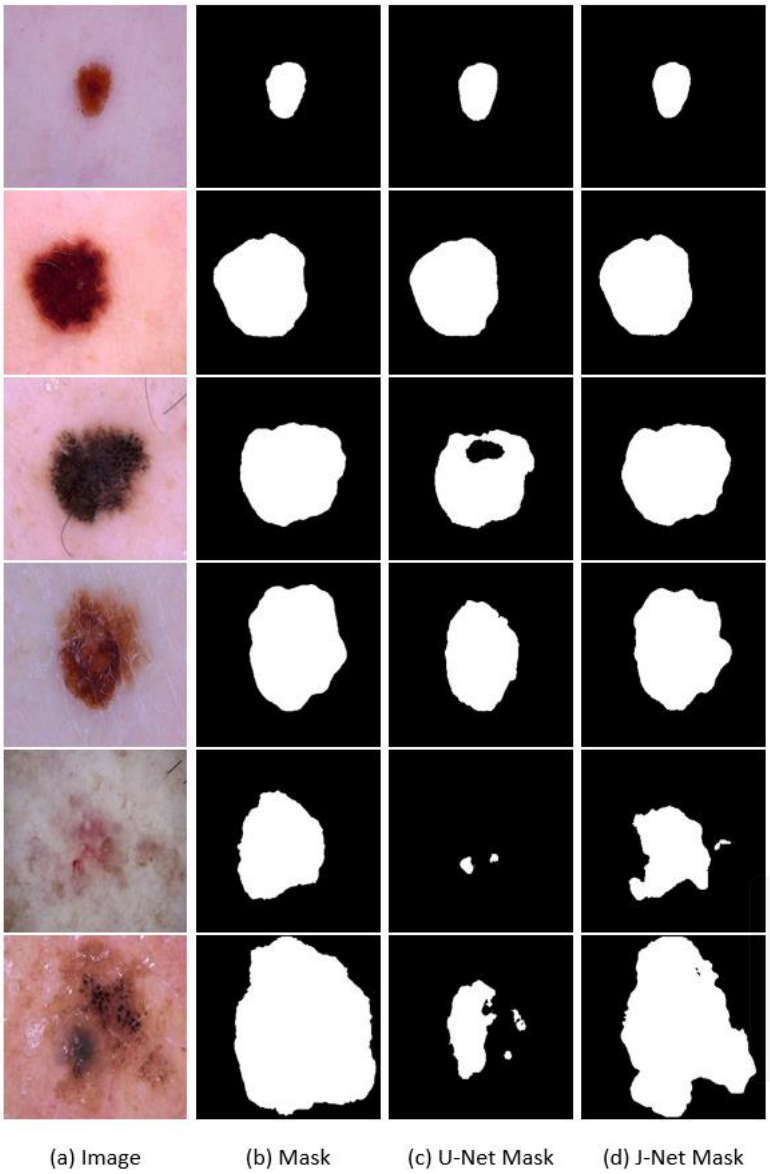

Figure 4-4: From left to right: skin cancer images, corresponding original segmentation masks, U-Net generated masks, and J-Net generated masks.

To compute the statistical significance, the models are trained over 30 times and the test set BCE loss is observed. The mean distribution of BCE loss for J-Net is lower than the U-Net in the segmentation task showing over 8% improvement.

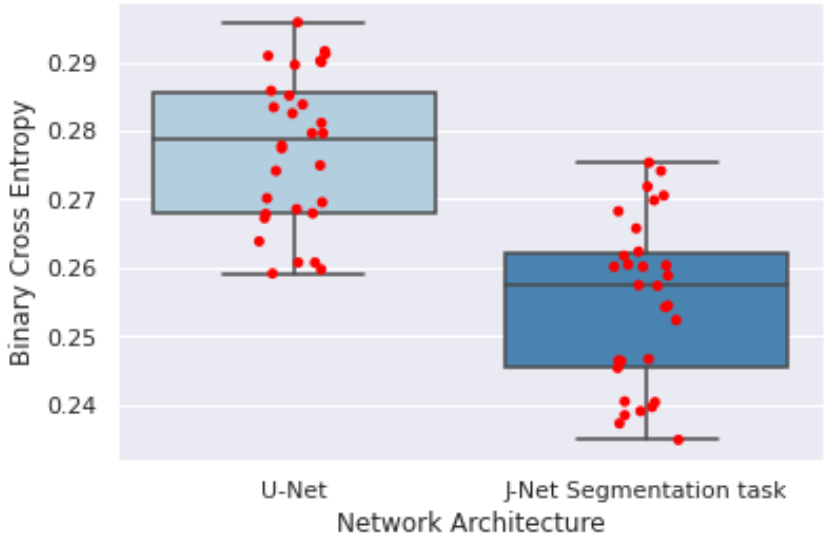

Figure 4-5: Boxplot results for 30 samples of BCE loss from lesion segmentation branch from U-Net vs J-Net classifier with $p < 0.05$.

Similarly, the J-Net classifier outperformed a simple binary classifier with the same architecture with over 5% improvement on the BCE loss. This experiment shows that adding additional information in training models helps them to adapt the hidden patterns from all perspectives.

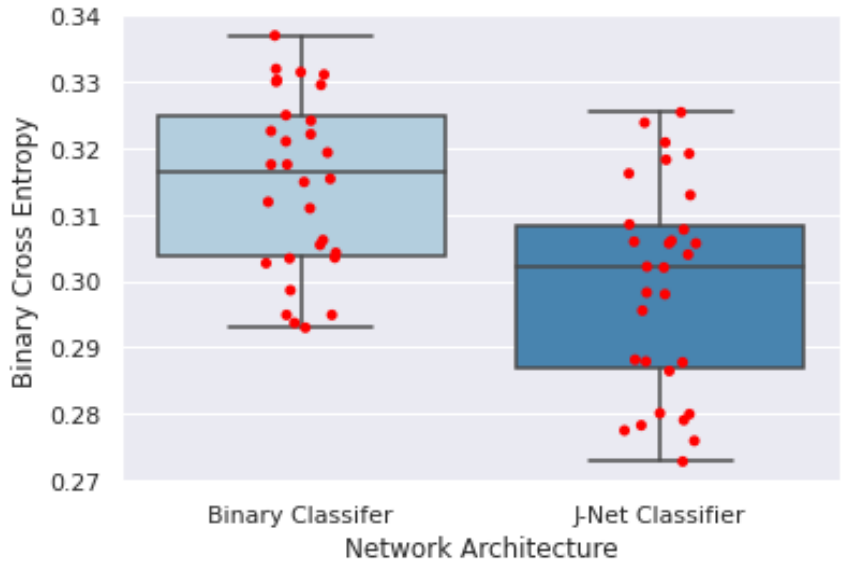

Figure 4-6: Boxplot results for 30 samples of BCE loss from binary image classifier vs J-Net classifier with $p < 0.05$.

Both the models are evaluated with performance metrics for the lesion segmentation task. Performance metrics like dice score, intersection over union (IOU), accuracy, precision, recall, F1, and binary cross entropy loss scores are recorded with varying training datasets between the models.

Table 4-1: Segmentation task metrics results between U-Net and J-Net with varying training data size from 200 to 1200.

size	U-Net Segmentation results				J-Net Segmentation results			
	Precision	Accuracy	Dice	IOU	Precision	Accuracy	Dice	IOU
1200	0.8200	0.9206	0.9254	0.7608	0.8719	0.9589	0.9637	0.8078
1100	0.8100	0.9113	0.9175	0.7534	0.8677	0.9376	0.9586	0.8022
1000	0.8000	0.9078	0.9107	0.7485	0.8446	0.9310	0.9478	0.7969
900	0.7981	0.8914	0.9013	0.7471	0.8381	0.9274	0.9279	0.7841
800	0.7959	0.8676	0.8971	0.7366	0.8282	0.9113	0.9147	0.7713
700	0.7801	0.8459	0.8864	0.7357	0.8196	0.8998	0.9063	0.7671
600	0.7677	0.8305	0.8849	0.7301	0.8082	0.8895	0.8872	0.7505
500	0.7612	0.8282	0.8705	0.7294	0.8031	0.8783	0.8838	0.7443
400	0.7582	0.8207	0.8699	0.7273	0.7996	0.8740	0.8757	0.7386
300	0.7503	0.8056	0.8623	0.7238	0.7925	0.8661	0.8685	0.7321
200	0.7497	0.7993	0.8571	0.7206	0.7844	0.8604	0.8602	0.7284

The classification results between J-Net and a binary classification model are drawn for comparison. The binary classifier has the same architecture as the J-Net classifier to ensure the fairness of the results. The outcomes of the Y-Net classifier outperformed the binary classifier with better precision and Recall. Thus, improved in categorizing the images into mild and severe melanocytic nevi classes. This study can be easily extended to other domains with limited image data but with additional attributes about the images.

Table 4-2: Classification model performance metrics between Binary classifier and J-Net classifier with varying training data size from 200 to 1200.

size	Binary Image Classifier results				J-Net Classifier results			
	Precision	Accuracy	Recall	F1	Precision	Accuracy	Recall	F1
1200	0.74	0.90	0.73	0.72	0.78	0.93	0.76	0.76
1100	0.74	0.89	0.72	0.72	0.77	0.92	0.76	0.76
1000	0.73	0.89	0.72	0.72	0.76	0.91	0.76	0.76
900	0.72	0.88	0.70	0.70	0.76	0.90	0.74	0.74
800	0.70	0.86	0.68	0.68	0.74	0.89	0.72	0.72
700	0.68	0.85	0.66	0.66	0.72	0.88	0.70	0.70
600	0.69	0.85	0.68	0.68	0.72	0.87	0.70	0.70
500	0.68	0.84	0.68	0.68	0.71	0.85	0.68	0.68
400	0.68	0.82	0.66	0.66	0.70	0.84	0.68	0.68
300	0.68	0.81	0.66	0.72	0.69	0.82	0.68	0.67
200	0.66	0.80	0.65	0.72	0.68	0.82	0.66	0.66

4.4 Discussion

In this study, we present a novel study that analyzes and compares two methods of applying two independent models separately and one joint model to conduct lesion segmentation and classification tasks, as well as lesion segmentation and classification performance changes of these two methods. From the study results, we observe the following two interesting aspects:

1. The J-Net joint model yields higher lesion segmentation and classification accuracy than using two independent single models even with small datasets (as shown in Tables 4-1 and 4-2). This can be explained due to the features learned from two integrated branches of the J-Net model, which helps improve model performance in both lesion segmentation and classification. Thus, combining two tasks of lesion segmentation and classification into one model allows for better integrating image features and the context learned from different perspectives.
2. This study demonstrates a clear trend of performance increase of the deep learning models including U-Net and joint J-Net models as the size of training and testing dataset increases (from 200 to 1200 in this study). Observation of the performance increase trend clearly indicates that increasing training dataset size and diversity plays a very important role in developing deep leaning models using medical images. Thus, in future research, more effort should be added to increase training dataset size by either collecting more clinical images or developing more effective algorithms or models to produce more clinically relevant synthetic images.

Although this is a quite unique study to address two important issues in optimally applying deep learning models in medical image research, we recognize that this is a quite preliminary study with several limitations including using only 2D images of skin cancer and difficulty to obtain accurate lesion segmentation masks due to inter-reader variability. Thus, more studies are needed to investigate more robust approaches to optimally develop and apply deep learning technologies and models for different medical imaging application tasks in the future.

5 Comparison of Performance in Breast Lesions Classification Using Radiomics and Deep Transfer Learning: An Assessment Study

5.1 Two Types of Feature Engineering

Full-field digital mammography (FFDM) is the most common and widely accepted clinical imaging modality for breast cancer screening in the general population. However, FFDM has a relatively lower sensitivity and specificity due to two-dimensional projection imaging. Thus, it is challenging to develop computer-aided detection and diagnosis (CAD) to assist radiologists in detecting suspicious lesions and classifying between malignant and benign lesions. Currently, computer-aided detection (CADe) schemes have been routinely implemented in the clinical practice, while computer-aided diagnosis (CADx) schemes have not been accepted in clinical practice. In previous CAD studies, two technologies have been widely used to extract and compute image features for lesion classification.

First, traditional feature engineering to capture radiomic information is popular and well accepted in developing CAD schemes of medical images. Based on the radiomics concept, CAD schemes can extract a vast number of handcrafted features specific to understand the underlying phenomenon of suspicious breast lesions. These radiomic features can be obtained from a wide range of attributes covering shape, density, texture patterns, frequency domain features etc. However, this higher number of initial feature dimensions comes with redundant information between features due to high correlations. Thus, it is important to take precautionary measures to reduce feature numbers or dimensionality. Optimal features can be obtained from either feature selection or reduction techniques. Additionally, the radiomics approach often faces the challenge of accurately segmenting subtle lesions if needed before feature engineering of local regions.

Second, in recent years, interest in extracting automated features using deep transfer learning is emerging. Transfer learning exploits the phenomenon of learning global features independent of image types to initialize the network weights on larger and more commonly available images. Then, the pre-trained network can be fine-tuned using a small medical image dataset related to the specific application task. However, medical professionals do not readily trust these "black-box" type, image-in and prediction-out schemes for medical image analysis.

Since in previous studies, CAD schemes are separately developed using either handcrafted radiomics features or deep transfer learning model [70]–[72] generated automated features using different and relatively small image datasets. Thus, it is very difficult to compare the performance of these two types of image features to achieve better performance. As a result, the advantages and/or potential limitations of CAD schemes trained using radiomics and automated features have not been well investigated to date. In order to address this issue, we conduct a new study to explore the association/correlation between the traditional radiomics feature-based CADs and a deep learning framework-based CAD scheme in classifying between malignant and benign breast lesions using a relatively large and diverse image dataset. Additionally, we also investigate whether integration of these two types of features further improves performance in lesion classification.

5.2 Materials and methods

5.2.1 FFDM Image Dataset

A fully anonymous and retrospective database consisting of full-field digital mammograms (FFDMs) was assembled for this study. The dataset is heterogeneous and consists of 2,778 FFDM images from craniocaudal (CC) and mediolateral oblique (MLO) views. The center location

belonging to suspicious lesion (soft tissue mass) in each image was marked by the radiologist. Based on biopsy results, these images depict 1,452 malignant and 1,326 benign masses.

The study primarily consists of two main phases, (i) a traditional image analysis phase with details involving handcrafted features (ii) a deep learning architecture adjusted and fine tuned for generating probabilities to classify between benign and malignant classes. More details regarding these two phases, along with model evaluation settings, are explained in the following sections.

5.2.2 Image Processing and Traditional Feature Engineering

During the traditional image processing phase, we first examined all the images to identify the ideal size of a rectangular window centered around the radiologists marking enclosing the lesion region. We observed the optimal window size to be 150×150, covering all types of lesions in the dataset. Then, we cropped the fixed-size image patches centered with the reference markings for each case. Necessary steps were taken to zero pad the edges or corners if the central region is along the boundary. Additionally, a relatively small subsample of cases consisting of chest wall regions within the patch was automatically segmented out. Then, we performed an adaptive thresholding-based segmentation with the seed selected at the center. The segmentation results are generally satisfactory, and only a small subset (<5%) needed a manual adjustment of the segmentation boundary.

Next, a total of 235 traditional handcrafted image features covering a variety of radiomic information representing lesion characteristics such as shape, density, boundary contrast, texture patterns, and wavelets were computed. The lesion-specific features explain local patterns like shape and density distribution within and around the boundary region. In contrast, global image features capture the total image patch’s texture, density patterns, and frequency domain

information. More detailed information regarding these features can be found in our previous studies [1, 2].

5.2.3 Deep Learning Framework Settings

We used the popular image classification architecture of pre-trained residual net architecture (ResNet50) for the deep learning phase with weights tuned for the ImageNet dataset consisting of 1,000 classes. The final fully connected (*FC*) layer used for prediction was adjusted to categorize two classes (benign or malignant lesions). Then we feed image patches of size 150×150 into the deep learning architecture and required transformations such as resizing (224×224×3: Height × Width × Depth)-normalization of the mean and standard deviation of each channel were performed on the fly. We used the same grayscale FFDM image patch repeated for the three channels for depth. Additionally, a minimal augmentation step (involving random centered crop, random horizontal, and random vertical flip with p=0.5) was added to introduce slight variation of a sample image for different epochs during the training phase. Due to the nature of medical images, a simple feature extractor type training involving freezing of all unchanged layers and updating only the weights and biases of the modified last FC layer did not yield good results. Thus, in this study, we optimized the weights of all layers during training.

Given the limitation of our dataset size relative to the computer vision field, we maximize the training and consider the time required for this network-tuning; currently, we used 10-fold cross-validation (CV). During each fold, the data is split randomly into training (90%) and testing (10%) without repetition between them, and each sample case is only used once in the test phase. We investigated various batch sizes (i.e., 4, 8, 16, etc.) and found that a batch size of 4 works well for

our analysis. Additionally, we selected the Adam optimizer with an initial learning rate (*lr*) of 1e-4 at the beginning of each cross fold. We updated the learning rate scheduler with an exponential decay function with a gamma value of 0.4 after each epoch. After each epoch, the network was evaluated to monitor training and validation loss during the training process, thereby deciding the stopping criterion. We noticed that by ten epochs, the network is saturated, and any further training resulted in overfitting. Thus, we only trained the network for ten epochs during each cross fold. During the validation phase, the network is loaded in evaluation mode, and a forward pass of data is done to collect both classification labels and probabilities. In summary, we used a 10-fold CV with ten epochs per each fold; at the end of each training fold, the test data was evaluated on the network to record both classification labels and the associated probabilities.

5.2.4 Model Building and Performance Evaluation:

We build and test several models to classify suspicious breast masses into two classes. Specifically, we investigated: (i) using only standard radiomic features, (ii) using probability score from ResNet50, (iii) integrated models with a combination of radiomic and ResNet50 models. In Model I, the initial feature dimension of size 235 is reduced using PCA with a variance rate of 0.99, and then an SVM classifier was implemented. In Model II, a simple classification based on a prediction probability of ResNet50 was performed. For Model III, multiple combinations of the above two were conducted using the output scores of Models I and II, including Model III.1 using two scores from Model I and II considered as features to build an SVM classifier, Model III.2 using a simple weighted average of classification scores generated by Models I and II, and Model III.3 using a minimum score of Models I and II, and Model III.4 using a maximum score of models I and II. Additionally, we also investigated the classification performance of three subgroups of traditional radiomic features (a. shape + density, b. wavelets, c. texture groups) with the integration of PCA

into their respective SVM classification learner. The classification scores of each model were named using '*S*' followed by the subscript of the model number. For instance, the Model I output score is S_1, and the output score of a weighted average model built using a combination of S_1 and S_2 is termed as $S_{3.2}$.

To evaluate the performance of each model, we used two steps. First, a receiver operating characteristic curve (ROC) is constructed from the classification scores. The area under the ROC curve (AUC) is computed and used as an index to evaluate and compare the performance of each model to classify between two classes. Second, we apply an operating threshold on the classification scores (T = 0.5) to divide all testing cases into two classes (score ≤ 0.5: 'Benign'; score > 0.5: 'Malignant'). Figure 5-1 shows a detailed flow chart explaining each step of the proposed CAD scheme.

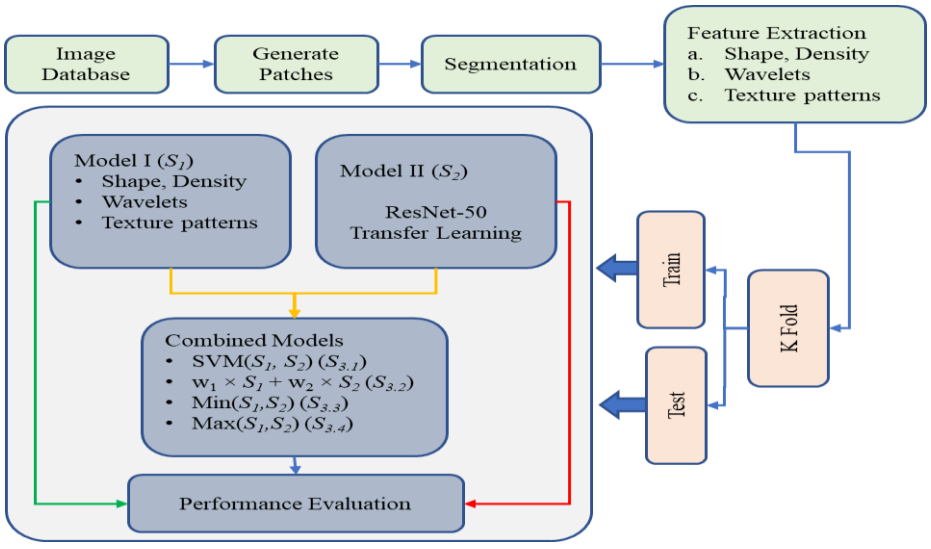

Figure 5-1: A detailed flow diagram of each step of the proposed CAD scheme.

5.3 Model Performances on FFDM Dataset

Figure 5-2 shows sample images in the database with an overlay of the segmentation results. The cases with segmentation overlay with red or green color are malignant or benign cases, respectively. Additionally, in some image patches, we can also notice that zero paddings are performed whenever needed (lesion at the edge or corner inside the original image). We can also notice that the density distribution of lesions consists of both solid and diffused samples. The diffused or hidden lesions segmentation is challenging to segment and/or analyze and may not represent the underlying lesion image marker.

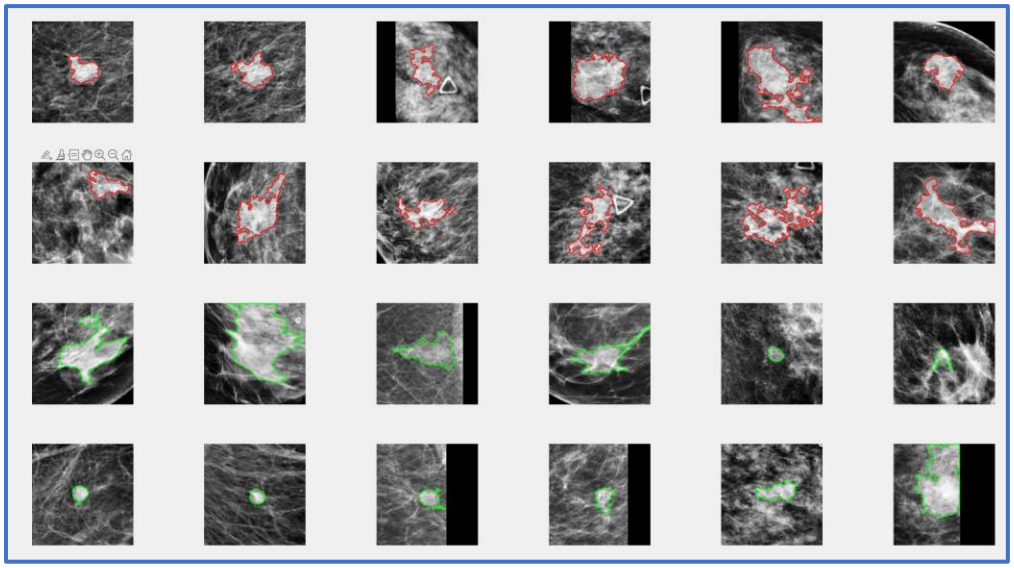

Figure 5-2: Sample case image patches with segmentation overlay (Red: Malignant; Green: Benign).

We first performed an independent analysis of each subgroup in Model I during the performance evaluation stage before the models. These models were also evaluated using PCA integration with an SVM classifier during each cross fold. The first model was built using a subgroup of features, including shape and density-related features. It included a total of 50 features, and the distribution of ACCs was 65.68±0.02. Similarly, the models built using wavelet and texture pattern features

independently resulted in an ACC distribution of 64.39±0.04 and 61.94±0.02, respectively. These traditional radiomic-based models' performance was the lowest as expected as we were only observing the classification capabilities separately. Whereas, when a combined model of these three subgroups was performed (Model I), we observed an increase in the performance metrics (ACC, AUC) as shown in Figure 5-3, indicating that the combination of these types adds new information for classification model to learn new information. Next, the ResNet50 (Model II) performed significantly better than the Model I in terms of both ACC (77.31±2.65) and AUC (0.85±0.02). The trendline depicting the change in the improvement of performance distribution in terms of ACC for each fold per epoch using the ResNet50 is shown in Figure 5-3.

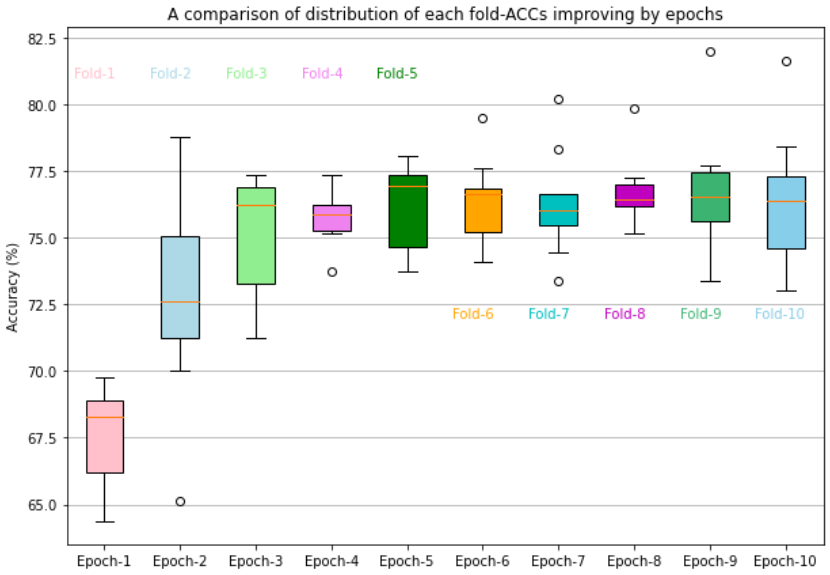

Figure 5-3: Trendline depicting the change in ACC distribution for each fold per epoch using ResNet50.

Next, we used four different combinations to observe the performance improvement combining Models I and II scores. In Models III.1 and III.2 built using SVM and weighted averages of S_1 and S_2, we noticed that the performance metrics are very similar to that of ResNet50. This indicates that both traditional and deep learning features converge at the end towards classification

prediction and have a high correlation. Additionally, a negative effect on performance was observed when using either min- or max-based simple classification models. A more detailed comparison of the distribution of performance metrics for each model is shown in Figure 5-4. In terms of both ACC and AUC from these results, we clearly notice that ResNet50 or a combination of ResNet50 with traditional radiomic features yield a similar performance with a high association.

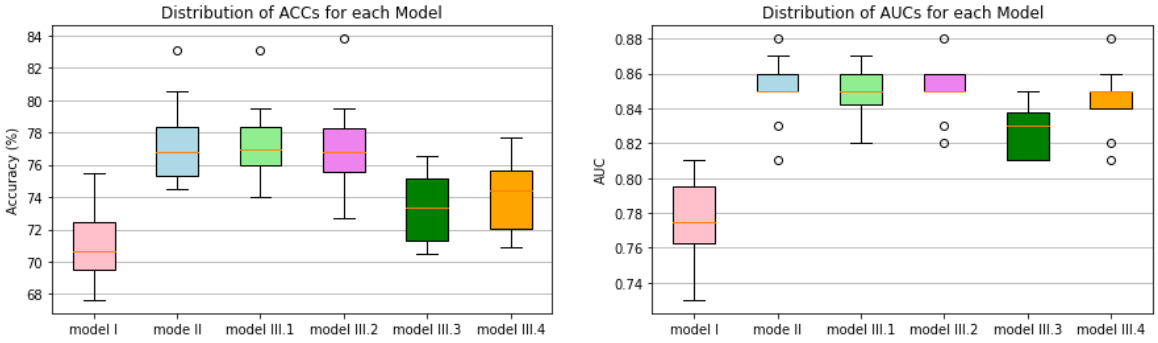

Figure 5-4: Comparison between the distribution of (a) accuracies and (b) AUCs for each model.

Table 5-1: Summary of details for each model, including feature description and performance metrics.

Model	Feature description	AUC	ACC (%)
Model I (S_1)	a. shape + density, b. wavelets, c. texture groups	0.77 ± 0.02	71.23 ± 2.44
Model II (S_2)	classification probability of ResNet50	0.85 ± 0.02	77.31 ± 2.65
Model III.1 ($S_{3.1}$)	SVM (S_1, S_2)	0.85 ± 0.01	77.42 ± 2.47
Model III.2 ($S_{3.2}$)	$W_1 \times S_1 + W_2 \times S_2$	0.85 ± 0.01	77.31 ± 2.83
Model III.3 ($S_{3.3}$)	Min (S_1, S_2)	0.83 ± 0.02	73.35 ± 2.17
Model III.4 ($S_{3.4}$)	Max (S_1, S_2)	0.85 ± 0.02	74.07 ± 2.24

5.4 Discussion

In this study, we investigated the association between the traditional radiomics-based CAD features and the deep learning framework (ResNet50) in the classification of breast masses. This study generates several new and interesting observations, namely, first, the performance of subgroups of radiomic features is low when evaluated separately, whereas, when combined, new information from each subgroup contributes additional information for the classification learner to improve the performance (ACC) significantly from a range for individual models (65.68±0.02, 64.39±0.04, 61.94±-0.02) to the combined Model I (71.23±2.44).

Second, training a complex deep learning framework (ResNet) by freezing all input layers is not ideal given the contrast between ImageNet data and complex structures in the breast region. However, a significant improvement in results compared to Model I is achieved by retraining a

transfer learning model to update weights of all the layers in the network. Initializing the deep learning framework with weights from pre-trained ImageNet and customizing for a binary classification task (i.e., classifying between malignant and benign breast lesions in this study) works well. This step of careful customization and training all the layers for certain epochs is essential for optimally applying the deep transfer learning network to learn the parameters used in CAD of medical images.

Third, when we combine the scores from Models I and II, we observe no significant change in improvement of the performance as compared to Model II. Nevertheless, this supports the theory that both optimized radiomic features and deep learning features have a high degree of correlation. This level of closeness/association between two classification scores represents that even though no specific attention is taken in the deep learning framework after careful selection of features, they still capture the underlying phenomenon of breast images similar to that of radiomic features. Thus, even a simple weighted average base classifier model built using a combination of scores S_1 and S_2 yielded a consistent performance similar to that of ResNet50.

Despite the encouraging observations in our results, we recognize some limitations in our study. First, even though we used a wide range of radiomic features (shape, density, texture, wavelets) for our Model I, there are numerous more combinations to analyze the traditional feature-based model. Second, we only used the standard ResNet50 by modifying the last *FC* layer to examine and utilize the full potential of transfer learning. We need to validate this phenomenon on other state-of-the-art deep learning frameworks in the future. Third, we were limited to using 10-fold cross-validation given the constraints; in the future, we will try to use either more folds or obtain a separate new test dataset to validate the classification performance of these transfer learning models.

In conclusion, this study observes and compares traditional radiomic feature-based CADs and deep learning framework-based CAD to classify breast masses. Additionally, we also observed a high degree of correlation between the classification scores of two types of CAD models, representing that both preserve/capture similar information irrespective of the discrepancies in both approaches. Thus, although deep transfer learning is widely considered a "black-box" type study with a high degree of difficulty for human users to understand its learning or decision-making logic or reasoning, the automated features provide high discriminatory information or power compared to traditional radiomics features. A more comprehensive analysis covering both radiomic and deep learning architectures needs to be investigated to validate these observations.

6 Conclusions and Future Work

In recent decades, there has been a growing interest in the development of Computer-Aided Diagnosis (CAD) systems for medical images, both in commercial companies and research institutions. These CAD systems have found applications in clinical research and practice, often serving as “second readers” to aid radiologists in image interpretation. However, their adoption in clinical practice remains limited, necessitating further development efforts.

To create robust and reliable CAD schemes, several stages of supervision are crucial:

1. Image Quality Enhancement: The initial stage involves eliminating artifacts and improving image quality to ensure accurate analysis.
2. Region of Interest Detection and Segmentation: CAD systems need to accurately identify and delineate the areas of interest within the medical images.
3. Radiomic Feature Extraction: Radiomic features are computed and optimized to capture valuable information from the images.
4. Machine Learning Model Fine-Tuning: Parameters of the machine learning models are fine-tuned to align with the specific medical application objectives.

Recent advancements have accelerated the development and implementation of CAD systems in medicine. These catalysts include:

1. Improved Imaging Technologies: Enhanced imaging technologies provide higher-quality input for CAD systems.
2. Computational Speed: Growing computational processing speeds enable more efficient CAD algorithms.

3. Radiomics Concept: The emergence of radiomics, which connects image features with genomic and radiologic markers, has expanded the possibilities for CAD.
4. Deep Learning: The application of deep learning architectures has revolutionized medical image analysis.

Research Interest in Medical Imaging Informatics: The increasing focus on machine learning in medical imaging research has further propelled CAD development. Despite the promising results achieved with CAD systems in medicine, this field is still emerging and requires greater research involvement across various domains. Many clinical studies involving medical image interpretation can benefit from identifying novel radiographic image markers that capture clinically relevant patterns. Manual interpretation of these images often suffers from high variability between different readers and is time-consuming and inconsistent. The concept of radiomics has demonstrated its potential in assessing radiographic images, such as CT and MRI, by extracting features highly correlated with genomic biomarkers and disease prognosis prediction. Incorporating study-specific radiomic-engineered CAD systems alongside existing radiologist evaluation tools can serve as valuable secondary readers, addressing some of the existing limitations and enhancing the accuracy and efficiency of medical image interpretation.

In Chapter 2, we proposed, developed, and tested a new two-stage model in classifying retinal fundus images using transfer learning ResNet-50 models. Additionally, we also observed that like conventional machine learning or computer-aided detection schemes, performance of a deep-learning model also heavily depends on the content distribution of training datasets. Thus, it is often difficult to correctly identify or classify the difficult or subtle images using a deep-learning model that is trained or fine-tuned using a small dataset which does not contain or cannot sufficiently represent the difficult images (or outliers). In order to optimally address this challenge,

developing a two or multi-stage deep-learning approach has significant advantages. This is the primary contribution of this study to the medical imaging informatics or CAD field. We also recognize that many previous studies have highlighted the significance of integrating image processing techniques along with deep learning frameworks. Thus, in our future studies, we plan to explore this phenomenon using a more diverse dataset of retinal images and investigate application of integrated architectures that combine both traditional image processing and deep learning frameworks.

In Chapter 3, we demonstrated a new algorithm to generate synthetic retinal fundus images embedded with different types of diseased lesions. This study has several unique characteristics and interesting observations. First, unlike other existing synthetic image data generation methods (i.e., using a Monte Carlo simulation or a Generative Adversarial Network), which are quite difficult to design and computationally expensive because large numbers of algorithm or network parameters need to be chosen and optimized, our new method is much simpler and more computationally efficient. It uses a multi-stage approach to directly extract positive lesions, randomize distribution of lesion blobs, and then insert positive lesion blobs in randomly selected locations of negative images. In the medical imaging field, collecting large numbers of negative images is much easier than collecting positive images with manual annotation of diseased lesions. In this study, we significantly expand our dataset size from 54 positive and 60 negative images to 14,994 synthetic images (for Categories 1 and 2 diseases as shown in Table 3-1). Finally, we recognize that seamless insertion of lesions or other abnormality regions onto negative images using this simple algorithm has restrictions including that (1) the lesions must have clear boundary contours so that the lesions can be easily extracted, (2) the normal tissue background should also be relatively uniform. Retinal fundus images meet these two restrictions. Some other medical

images (i.e., liver tumors) can also meet these restrictions. Nonetheless, if lesions have fuzzy boundary embedded under heterogenous tissue background (i.e., breast tumors depicted on mammograms), this algorithm will not work "as is" and modifications will be needed. Despite such limitations, developing this new simple algorithm to generate synthetic images has its higher clinical impact at least for retinal fundus images that are acquired using a low-cost image examination method and widely used in clinics to screen, detect, and diagnose many common human diseases including eye diseases and diabetes. We will further test and validate this new algorithm and apply it to develop more accurate and robust DCNN models for different medical applications in the future.

In Chapter 4, we present a novel study that analyzes and compares two methods of applying two independent models separately and one joint model to conduct lesion segmentation and classification tasks, as well as lesion segmentation and classification performance changes of these two methods. From the study results, the J-Net joint model yields higher lesion segmentation and classification accuracy than using two independent single models even with small datasets (as shown in Tables 4-1 and 4-2). This can be explained due to the features learned from two integrated branches of the J-Net model, which helps improve model performance or accuracy in both lesion segmentation and classification. Thus, combining two tasks of lesion segmentation and classification into one model allows for better understanding and integrating image features and the context learned from different perspectives. Although this is a quite unique study to address two important issues in optimally applying deep learning models in medical image research, we recognize that this is a quite preliminary study with several limitations including using only 2D images of skin cancer and difficulty to obtain accurate lesion segmentation masks due to inter-reader variability. Thus, more studies are needed to investigate more robust approaches to

optimally develop and apply deep learning technologies and models for different medical imaging application tasks in the future.

In Chapter 5, we demonstrated a comparative study that successfully assesses and compares the two most popular types of CAD approaches using either traditional radiomics-based or deep learning such as CNN-based CAD models. We assembled a relatively large image dataset of nearly 2,800 mammograms to investigate the association/correlation between these two types of CAD schemes in evaluating their performance in classifying suspicious breast lesions. The study results show that both types of CAD schemes contain similar information in their ability to classify breast lesions. Additionally, the CNN-based CAD model performs significantly better than the traditional radiomics feature-based CAD model.

In conclusion, the culmination of the chapters in this dissertation represents a multifaceted exploration into advancing the field of Computer-Aided Diagnosis (CAD) for medical images. From the inception of robust CAD schemes, leveraging image quality enhancement, region of interest detection, radiomic feature extraction, to the incorporation of deep learning architectures and the innovative generation of synthetic images, each chapter contributes a vital piece to the evolving puzzle of medical imaging informatics. The collaborative efforts presented herein underscore the interdisciplinary nature of our approach, aiming to enhance the accuracy and efficiency of medical image interpretation. As we reflect on these contributions, we recognize the progress made, yet acknowledge the ongoing journey ahead. Future studies will delve deeper into integrated architectures, diverse datasets, and validation of novel algorithms, striving to push the boundaries of CAD development. The chapters collectively lay the groundwork for further exploration, and their impact extends beyond these pages, influencing the trajectory of research and practice in the dynamic landscape of medical imaging.

Bibliography

[1] M. C. B. Godoy *et al.*, "Benefit of computer-aided detection analysis for the detection of subsolid and solid lung nodules on thin- and thick-section CT," *American Journal of Roentgenology*, vol. 200, no. 1, pp. 74–83, Jan. 2013, doi: 10.2214/AJR.11.7532.

[2] J. J. Fenton *et al.*, "Effectiveness of computer-aided detection in community mammography practice," *J Natl Cancer Inst*, vol. 103, no. 15, pp. 1152–1161, Aug. 2011, doi: 10.1093/jnci/djr206.

[3] B. Sahiner *et al.*, "Effect of CAD on Radiologists' Detection of Lung Nodules on Thoracic CT Scans: Analysis of an Observer Performance Study by Nodule Size," *Acad Radiol*, vol. 16, no. 12, pp. 1518–1530, Dec. 2009, doi: 10.1016/j.acra.2009.08.006.

[4] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis," 2017, doi: 10.1146/annurev-bioeng-071516.

[5] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," *British Journal of Radiology*, vol. 78, no. SPEC. ISS. 2005. doi: 10.1259/bjr/82933343.

[6] S. Mudduluru, "Developing and Applying Hybrid Deep Learning Models for Computer-Aided Diagnosis of Medical Image Data", Accessed: Nov. 04, 2023. [Online]. Available: https://shareok.org/handle/11244/337603

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," Institute of Electrical and Electronics Engineers (IEEE), Mar. 2010, pp. 248–255. doi: 10.1109/cvpr.2009.5206848.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[10] S. K. R. Maryada *et al.*, "Applying a novel two-stage deep-learning model to improve accuracy in detecting retinal fundus images," SPIE-Intl Soc Optical Eng, Apr. 2022, p. 21. doi: 10.1117/12.2611565.

[11] S. Kiran Maryada *et al.*, "An EEcient Synthetic Data Generation Algorithm to Improve EEcacy of Deep Learning Models of Medical Images", doi: 10.21203/rs.3.rs-2416807/v1.

[12] S. Mudduluru, S. K. R. Maryada, W. L. Booker, D. F. Hougen, and B. Zheng, "Improving medical image segmentation and classification using a novel joint deep learning model," in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., SPIE, 2023, p. 124652H. doi: 10.1117/12.2654052.

[13] G. Danala *et al.*, "A Comparison of Computer-Aided Diagnosis Schemes Optimized Using Radiomics and Deep Transfer Learning Methods," *Bioengineering*, vol. 9, no. 6, Jun. 2022, doi: 10.3390/bioengineering9060256.

[14] D. L. Pham, C. Xu, and J. L. Prince, “CURRENT METHODS IN MEDICAL IMAGE SEGMENTATION 1,” 2000. [Online]. Available: www.annualreviews.org

[15] T. McInerney and D. Terzopoulos, “Deformable models in medical image analysis: a survey,” *Med Image Anal*, vol. 1, no. 2, pp. 91–108, Jun. 1996, doi: 10.1016/S1361-8415(96)80007-7.

[16] J. B. A. Maintz and M. A. Viergever, “A survey of medical image registration,” *Med Image Anal*, vol. 2, no. 1, pp. 1–36, Mar. 1998, doi: 10.1016/S1361-8415(01)80026-8.

[17] C. Li, R. Huang, Z. Ding, J. C. Gatenby, D. N. Metaxas, and J. C. Gore, “A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI,” *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2007–2016, 2011, doi: 10.1109/TIP.2011.2146190.

[18] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, *Hierarchical, Learning-based Automatic Liver Segmentation*.

[19] D. Shen, S. Moffat, S. M. Resnick, and C. Davatzikos, “Measuring Size and Shape of the Hippocampus in MR Images Using a Deformable Shape Model,” *Neuroimage*, vol. 15, no. 2, pp. 422–434, Feb. 2002, doi: 10.1006/NIMG.2001.0987.

[20] Y. Zheng *et al.*, “Automatic aorta segmentation and valve landmark detection in C-Arm CT for transcatheter aortic valve implantation,” *IEEE Trans Med Imaging*, vol. 31, no. 12, pp. 2307–2321, 2012, doi: 10.1109/TMI.2012.2216541.

[21] M. Betke, H. Hong, D. Thomas, C. Prince, and J. P. Ko, “Landmark detection in the chest and registration of lung surfaces with an application to nodule registration,” *Med Image Anal*, vol. 7, no. 3, pp. 265–281, Sep. 2003, doi: 10.1016/S1361-8415(03)00007-0.

[22] R. Bellotti *et al.*, “A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model,” *Med Phys*, vol. 34, no. 12, pp. 4901–4910, 2007, doi: 10.1118/1.2804720.

[23] L. A. Meinel, A. H. Stolpen, K. S. Berbaum, L. L. Fajardo, and J. M. Reinhardt, “Breast MRI lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system,” *Journal of Magnetic Resonance Imaging*, vol. 25, no. 1, pp. 89–95, 2007, doi: 10.1002/jmri.20794.

[24] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[25] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Adv Neural Inf Process Syst*, vol. 4, no. January, pp. 3104–3112, 2014.

[26] S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda, “Fundamental technologies in modern speech recognition,” *IEEE Signal Process Mag*, vol. 29, no. 6, pp. 16–17, 2012, doi: 10.1109/MSP.2012.2209906.

[27] P. Porwal *et al.*, “IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge,” *Med Image Anal*, vol. 59, p. 101561, Jan. 2020, doi: 10.1016/J.MEDIA.2019.101561.

[28] M. A. Pérez, B. B. Bruce, N. J. Newman, and V. Biousse, “The use of retinal photography in nonophthalmic settings and its potential for neurology,” *Neurologist*, vol. 18, no. 6, pp. 350–355, 2012, doi: 10.1097/NRL.0b013e318272f7d7.

[29] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[30] D. C. G. Binsheng Zhao, Yongqiang Tan, Daniel J Bell, Sarah E Marley, Pingzhen Guo, Helen Mann, Marietta L J Scott, Lawrence H Schwartz, “Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals,” *Eur J Radiol*, vol. 82, pp. 959–968, 2013, doi: https://doi.org/10.1016/j.ejrad.2013.02.018.

[31] G. Danala, S. K. R. Maryada, M. Heidari, B. Ray, M. Desai, and B. Zheng, “A new interactive visual-aided decision-making supporting tool to predict severity of acute ischemic stroke,” SPIE-Intl Soc Optical Eng, Feb. 2020, p. 66. doi: 10.1117/12.2549614.

[32] G. Danala *et al.*, “Applying Quantitative Radiographic Image Markers to Predict Clinical Complications After Aneurysmal Subarachnoid Hemorrhage: A Pilot Study,” *Ann Biomed Eng*, vol. 50, no. 4, pp. 413–425, Apr. 2022, doi: 10.1007/s10439-022-02926-z.

[33] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, “Cross-modality deep feature learning for brain tumor segmentation,” *Pattern Recognit*, vol. 110, p. 107562, Feb. 2021, doi: 10.1016/J.PATCOG.2020.107562.

[34] M. Heidari *et al.*, “Applying a Random Projection Algorithm to Optimize Machine Learning Model for Breast Lesion Classification,” *IEEE Trans Biomed Eng*, vol. 68, no. 9, pp. 2764–2775, Sep. 2021, doi: 10.1109/TBME.2021.3054248.

[35] G. Danala *et al.*, “Developing new quantitative CT image markers to predict prognosis of acute ischemic stroke patients,” *J Xray Sci Technol*, vol. 30, no. 3, pp. 459–475, Apr. 2022, doi: 10.3233/XST-221138.

[36] G. Danala *et al.*, “A Comparison of Computer-Aided Diagnosis Schemes Optimized Using Radiomics and Deep Transfer Learning Methods,” 2022, doi: 10.20944/preprints202206.0112.v1.

[37] G. Danala *et al.*, “Developing interactive computer-aided detection tools to support translational clinical research,” SPIE-Intl Soc Optical Eng, Mar. 2022, p. 12. doi: 10.1117/12.2607273.

[38] M. R. Canales-Fiscal and J. G. Tamez-Peña, “Glaucoma classification using a morphological-convolutional neural network trained with extreme learning machine,” in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., SPIE, 2023, p. 124652F. doi: 10.1117/12.2654025.

[39] Fsbi. Robert M.NishikawaPhD, FAAPM, “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, doi: https://doi.org/10.1016/j.acra.2014.05.018.

[40] X. Chen *et al.*, “Recent advances and clinical applications of deep learning in medical image analysis,” *Med Image Anal*, vol. 79, p. 4, 2022, doi: 10.1016/j.media.2022.1024.

[41] M. A. Jones, R. Faiz, Y. Qiu, and B. Zheng, “Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features,” *Phys Med Biol*, vol. 67, no. 5, Mar. 2022, doi: 10.1088/1361-6560/AC5297.

[42] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.

[43] H. C. Shin *et al.*, “Medical image synthesis for data augmentation and anonymization using generative adversarial networks,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11037 LNCS, pp. 1–11, 2018, doi: 10.1007/978-3-030-00536-8_1.

[44] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, “Chest pathology detection using deep learning with non-medical training,” *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 294–297, 2015, [Online]. Available: https://www.semanticscholar.org/paper/Chest-pathology-detection-using-deep-learning-with-Bar-Diamant/5cc51fb6ecadc853cb4017a43fb644ad1b852bc1

[45] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” *Nat Biomed Eng*, vol. 5, no. 6, pp. 493–497, 2021, doi: 10.1038/s41551-021-00751-8.

[46] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, pp. 321–331, 2018, doi: 10.1016/j.neucom.2018.09.013.

[47] F. Mahmood *et al.*, “Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images,” *IEEE Trans Med Imaging*, vol. 39, no. 11, pp. 3257–3267, 2020, doi: 10.1109/TMI.2019.2927182.

[48] F. Mahmood, R. Chen, and N. J. Durr, “Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training,” *IEEE Trans Med Imaging*, vol. 37, no. 12, pp. 2572–2581, 2018, doi: 10.1109/TMI.2018.2842767.

[49] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, “CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection,” *IEEE Access*, vol. 8, pp. 91916–91923, 2020, doi: 10.1109/ACCESS.2020.2994762.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.00567

[51] G. Danala, S. K. R. Maryada, H. Pham, W. Islam, M. Jones, and B. Zheng, “Comparison of performance in breast lesions classification using radiomics and deep transfer learning: an assessment study,” SPIE-Intl Soc Optical Eng, Feb. 2022, p. 35. doi: 10.1117/12.2611886.

[52] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification,” Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.02385

[53] B. Bhattarai, S. Baek, R. Bodur, and T.-K. Kim, “Sampling Strategies for GAN Synthetic Data,” Sep. 2019, doi: 10.1109/ICASSP40776.2020.9054677.

[54] G. Danala *et al.*, “Classification of Breast Masses Using a Computer-Aided Diagnosis Scheme of Contrast Enhanced Digital Mammograms,” *Ann Biomed Eng*, vol. 46, no. 9, pp. 1419–1431, Sep. 2018, doi: 10.1007/s10439-018-2044-4.

[55] B. Xiao *et al.*, “PAM-DenseNet: A Deep Convolutional Neural Network for Computer-Aided COVID-19 Diagnosis,” *IEEE Trans Cybern*, vol. 52, no. 11, pp. 12163–12174, Nov. 2022, doi: 10.1109/TCYB.2020.3042837.

[56] D. Zhang *et al.*, “Exploring Task Structure for Brain Tumor Segmentation from Multi-Modality MR Images,” *IEEE Transactions on Image Processing*, vol. 29, pp. 9032–9043, 2020, doi: 10.1109/TIP.2020.3023609.

[57] B. Xiao *et al.*, “Follow the Sound of Children’s Heart: A Deep-Learning-Based Computer-Aided Pediatric CHDs Diagnosis System,” *IEEE Internet Things J*, vol. 7, no. 3, pp. 1994–2004, Mar. 2020, doi: 10.1109/JIOT.2019.2961132.

[58] E. Chung, W. T. Leung, S. M. Pun, and Z. Zhang, “A multi-stage deep learning based algorithm for multiscale model reduction,” *J Comput Appl Math*, vol. 394, pp. 1–21, 2021, doi: 10.1016/j.cam.2021.113506.

[59] L. He *et al.*, “A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants,” *Sci Rep*, vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-71914-x.

[60] H. C. Shin *et al.*, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.

[61] A. V. Varadarajan *et al.*, “Deep learning for predicting refractive error from retinal fundus images,” *Invest Ophthalmol Vis Sci*, vol. 59, no. 7, pp. 2861–2868, 2018, doi: 10.1167/iovs.18-23887.

[62] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, Accessed: Mar. 02, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556

[63] S. K. Zhou *et al.*, “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,” Aug. 2020, doi: 10.1109/JPROC.2021.3054390.

[64] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med Image Anal*, vol. 42, pp. 60–88, 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[65] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *CoRR*, vol. abs/1512.03385, 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[66] M. Hussain, J. J. Bird, and D. R. Faria, “A study on CNN transfer learning for image classification,” *Advances in Intelligent Systems and Computing*, vol. 840, no. October, pp. 191–202, 2019, doi: 10.1007/978-3-319-97982-3_16.

[67] Z. Yan, Y. Zhan, S. Zhang, D. Metaxas, and X. S. Zhou, *Multi-Instance Multi-Stage Deep Learning for Medical Image Recognition,* 1st ed. Elsevier Inc., 2017. doi: 10.1016/B978-0-12-810408-8.00006-7.

[68] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[69] S. Mehta, E. Mercan, J. Bartlett, D. Weave, J. G. Elmore, and L. Shapiro, "Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.01313

[70] H. Pham *et al.*, "Identifying an optimal machine learning generated image marker to predict survival of gastric cancer patients," SPIE-Intl Soc Optical Eng, Feb. 2022, p. 79. doi: 10.1117/12.2611788.

[71] W. Islam, M. Jones, R. Faiz, N. Sadeghipour, Y. Qiu, and B. Zheng, "Improving Performance of Breast Lesion Classification Using a ResNet50 Model Optimized with a Novel Attention Mechanism," *Tomography*, vol. 8, no. 5, pp. 2411–2425, Oct. 2022, doi: 10.3390/tomography8050200.

[72] W. Islam, G. Danala, H. Pham, and B. Zheng, "Improving the performance of computer-aided classification of breast lesions using a new feature fusion method," SPIE-Intl Soc Optical Eng, Apr. 2022, p. 4. doi: 10.1117/12.2611841.