UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DEVELOPING AND APPLYING HYBRID DEEP LEARNING MODELS FOR COMPUTER-AIDED DIAGNOSIS OF MEDICAL IMAGE DATA

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

SANJANA MUDDULURU
Norman, Oklahoma
2023

DEVELOPING AND APPLYING HYBRID DEEP LEARNING MODELS FOR COMPUTER-AIDED DIAGNOSIS OF MEDICAL IMAGE DATA

A DISSERTATION APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Dean Frederick Hougen, chair

Dr. Sridhar Radhakrishnan

Dr. Chongle Pan

Dr. Charles Darren Nicholson

Acknowledgements

I want to begin by expressing my heartfelt gratitude to my advisor, Dr. Dean Hougen, for his guidance, support, and encouragement throughout my Ph.D. journey. His expertise, insights, and patience have been invaluable in shaping my research and enabling me to achieve my goals. I will always be grateful for his mentorship and friendship.

I also sincerely thank my research committee members, Dr. Sridhar Radhakrishnan, Dr. Charles Nicholson, and Dr. Chongle Pan. I have been fortunate to have such accomplished and supportive scholars on my committee, and I am honored to have benefited from their mentorship.

To my extended family, friends, and colleagues: you put up with me being distracted and missing many events. I am forever grateful for your patience and understanding. I hope to have time now to reconnect with each of you.

Finally, to my amazing parents and sister: your love and understanding helped me through the dark times. Without you believing in me, I would never have made it. It is time to celebrate; you earned this degree along with me.

Abstract

Although deep learning models have been widely used in medical imaging research to perform lesion segmentation and classification tasks, several challenges remain to applying deep learning models and improving model performance optimally. This dissertation highlights a new novel joint deep learning model to achieve both tasks simultaneously. Specifically, a novel J-Net (joint model) includes a two-way CNN architecture combining a U-Net model with an image classification model.

As a first demonstration of J-Net architecture, a skin cancer dataset with 1200 images, annotated lesion masks, and associated ground truth of 'mild' and 'severe' melanoma nevi status is used. The performance of the new joint model is compared with two independent models to perform lesion segmentation and classification separately. The results show that the performance of the J-Net is superior to the U-Net in image segmentation with improved accuracy by 8%. In addition, the J-Net image classification branch yields 5% better classification accuracy than a binary image classifier with the same model architecture. Moreover, in this dataset, 11 subsets are randomly generated from 200 to 1200 images with an incremental rate of 100. Each subset is then divided into training, validation, and testing groups using a ratio of 70:20:10, respectively. The study results show when training the models using data subsets of 200 to 1200 images, accuracy levels increase from 0.80 to 0.92 or 0.86 to 0.95 in lesion segmentation. The lesion classification rises from 0.80 to 0.90, or 0.82 to

0.93, using two single models and one joint J-Net model, respectively. Thus, this study demonstrates that applying this new J-Net joint model enables higher lesion segmentation and classification performance than two single independent models. Additionally, the J-Net model produces better accuracy with lower data volumes than separate models.

However, building a robust AI model requires a large and diverse dataset for training and validation. For melanoma nevi, such a dataset is readily available. Unfortunately, this is not the case for many medical diagnosis tasks, such as detecting lesions in retinal fundus images. While many fundus photos are available online, collecting them to create a clean, well-structured dataset is complex and manually intensive. Two multi-stage deep-learning methods are discussed to address the lack of large, diverse datasets.

Method 1: A two-stage deep-learning system is introduced to automatically identify clean retinal fundus images and delete images with severe artifacts. In two stages, two transfer-learning models based on the ResNet-50 architecture pre-trained using ImageNet data are built with increased threshold values on SoftMax to reduce false positives. The first stage classifier identifies 'easy' images, and the second stage classifier further identifies the remaining 'difficult' (or undetermined) images. Using the Google Search Engine, 1,227 retinal fundus images are retrieved. This two-stage deep-learning model yields a positive predictive value (PPV) of 98.56% for the target class compared to a single-stage model with a PPV of 95.74%. The two-stage model helps reduce false positives for the retinal fundus image class by two-thirds. The PPV over all classes increases from 91.9% to 96.6% without compromising the number of images classified by the model. The superior performance of this two-stage model indicates that building an optimal training dataset can play an essential role in increasing the performance of deep-learning models.

Method 2: An efficient multi-stage algorithm is introduced to generate synthetic medical image data by extracting annotated diseased regions and randomly projecting them onto

disease-free images. To test the feasibility of this new algorithm, the publicly available Indian Diabetic Retinopathy Image Dataset (IDRiD) is used. This dataset is comprised of the annotated fundus images acquired from 81 patients with two categories of diseases. Among them, 54 and 27 images are used for training and testing images, respectively. Using the proposed algorithm, synthetic data is generated by inserting extracted diseased lesions onto another set of 60 disease-free images, which results in 7,902 and 6,786 images for the two categories of diseases, respectively. Three transfer learning-based DCNN models (VGG16, ResNet50, and Inception-v3) are trained using original IDRiD images and synthetic datasets. When applied to the same test images, the model trained with the synthetic dataset outperformed the model trained using the original IDRiD dataset by 7.4% in disease classification.

In conclusion, this thesis discusses three methods to address the challenges of applying deep learning models to medical imaging. The first method involves the development of a new joint deep learning model, J-Net, which combines a U-Net model with an image classification model to achieve lesion segmentation and classification simultaneously. The J-Net model outperforms the individual models in accuracy with small datasets. The second method performs automatic image detection using a two-stage deep learning model to produce clean data. The third method involves developing multi-stage deep learning algorithms to generate synthetic medical image data, which can be used to overcome the lack of large, diverse datasets. These methods demonstrate that building enhanced training datasets can play a vital role in improving the performance of deep-learning models in medical imaging applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces medical image analysis and hybrid deep-learning models for disease diagnosis and prognosis prediction. It outlines the J-Net on which this dissertation is built. Apart from J-Net, two hybrid models are discussed. This chapter also introduces the general motivations and contributions of this dissertation.

1.1 Introduction to Medical Image Analysis

Medical image analysis is a rapidly growing field that uses computer algorithms and mathematical models to analyze medical images. With the development of medical imaging technologies, such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound, medical images have become essential tools in diagnosing and treating diseases. Medical image analysis enables healthcare professionals to extract useful information from these images, allowing for more accurate diagnoses and better treatment plans.

The analysis of medical images involves a wide range of techniques, including image processing, image segmentation, feature extraction, pattern recognition, machine learning,

and deep learning. These techniques are used to identify and quantify patterns in medical images, such as the presence of tumors or the progression of a disease. The results of medical image analysis can provide clinicians with valuable insights into the patient's condition, allowing them to make informed treatment decisions.

Medical Image Analysis has numerous medical applications, including diagnosis, treatment planning, and monitoring. It can diagnose various medical conditions, such as cancer, cardiovascular disease, and neurological disorders. Medical Image Analysis can also be used to plan surgical procedures, allowing surgeons to identify the optimal surgical approach and reduce the risk of complications. Additionally, it can be used to monitor the progress of a disease, allowing healthcare professionals to track the effectiveness of treatments over time.

1.1.1 Image Segmentation and Classification

The primary purpose of *Computer Aided-Diagnosis(CAD)* schemes (12; 4; 21) is to assist clinicians and researchers in a range of applications, from organ segmentation (51) to disease diagnosis (32). Applying CAD schemes of medical images has attracted broad research interest in the medical imaging informatics field. Most CAD schemes are developed using conventional machine learning methods or models during the initial research stages, often requiring domain expertise for image feature selection to build better CAD schemes. However, due to the difference between human vision and computer vision, effectively or optimally defining or selecting non-redundant and highly clinically relevant handcrafted medical image features is complicated and thus remains a significant challenge in developing CAD schemes.

However, recent studies have demonstrated that applying deep learning methods can successfully mitigate the shortcomings mentioned above and thus, it has attracted extensive research in the past decade (27; 20). Deep Learning has enabled advancements in many fields, including computer vision (17; 26).

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 various medical imaging datasets to diagnose many diseases (9; 10; 31). In addition, researchers have also tried to fuse or combine multiple deep-learning models to achieve higher efficiency in image processing (8; 9). These hybrid or joint models can process different data types and perform better than individually trained models. 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 biomedical imaging, analyzing image features from different models could help medical professionals understand the issue at hand.

Despite promising advantages of developing and applying deep learning models, limited access to data in medical domains hinders the application of deep learning models. Due to data regulations and market fragmentation, it can be challenging to fabricate large homogeneous datasets. Therefore, creating deep learning models that can work with datasets of limited size is still difficult. 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, a lack of knowledge can lead to overfitting in small datasets.

Providing additional information to the model by producing a segmentation mask of the input image can help to alleviate this overfitting problem. This mask typically has the same dimensionality as the input image, providing more information for the model to optimize. Hence, it is possible to improve model performance on smaller datasets by combining classification and segmentation tasks as a single joint network. Y-Net (31) 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 study aims to introduce a hybrid deep-learning model that simultaneously generates segmentation masks and diagnoses classification labels for skin cancer images. Our network provides a simple architecture that can quickly adapt to any dataset. Due to the straightforward nature of our network, smaller datasets with a few hundred data records can yield highly accurate results avoiding overfitting.

1.1.2 Image Detection

In medical image analysis, clinical data representation and extraction are the primary purposes and the premise of many complicated frameworks. Many voluntary and involuntary models for data extraction are used in medical image analysis (35; 30).

A retinal fundus image captures a color image of the inner surface of the eye and is instrumental in helping ophthalmologists detect eye disorders. In recent years, examining retinal fundus imagery has proven useful in other medical fields besides ophthalmology. Analyzing abnormalities in the optic nerve, retina, and choroid can detect early signs of neurological disorders [18]. The imagery also provides neurologists insight into underlying systemic diseases contributing to a patient's neurological condition. Thus, fundus photography can aid physicians in identifying severe ailments. A large, diverse, and clean dataset is required to develop accurate CAD models using artificial intelligence (AI) or deep machine learning (ML) for identifying underlying conditions. However, acquiring such datasets from private sources can be challenging due to regulations such as HIPAA.

One way to obtain a dataset is by using public data platforms such as the Google search engine. However, this approach can result in unwanted and noisy images that contain severe artifacts or unrelated items alongside useful fundus photography images when appropriate

keywords are used in the search engine. Selecting visually or manually cleaning fundus images from these images is tedious and time-consuming.

To enhance the effectiveness of deep learning models using relatively small image datasets, **??** aims to explore the feasibility and benefits of creating a new automated system to identify and choose useful retinal fundus images from all downloaded images without any human intervention. Each downloaded image is classified into one of three categories 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 desired category that is added to the database. In this study, achieving the highest possible positive predictive value (PPV) for Class 1 images is critical.

1.1.3 Sythetic Data Generation

Medical images are widely used to detect and diagnose diseases. As a result, imaging data accounts for approximately 90% of all clinically diagnostic data generated in healthcare systems

However, traditional machine learning classifiers remain challenging and not robust, so researchers are considering incorporating deep-learning or deep-transfer-learning models based on deep convolutional neural networks (DCNNs) into CAD systems. The main challenges faced by DCNN models are the need for much larger and more diverse datasets to reduce bias and achieve high levels of robustness. Two approaches commonly used to apply DCNN models in medical imaging are data augmentation and deep transfer learning. However, these two approaches can only partially reduce the impact of small image datasets. Researchers have investigated whether adding synthesized data generated from real medical image data can help overcome these issues and thus develop more robust DCNN models. One promising

solution to generate synthetic image data is GANs, but they are very complex machine-learning models. This study proposes a new and more computationally efficient algorithm to generate synthetic image data on a diverse image base. The algorithm extracts the annotated diseased region or lesion blobs, randomly redistributes these lesion blobs, and projects them onto negative (healthy) images to optimally achieve seamless insertion.

Currently, to apply DCNN models in the medical imaging field, two approaches are commonly used. The first one uses *data augmentation* techniques to increase the number of training data items and directly mitigate the paucity of annotated data. The second one is to apply deep transfer learning to mitigate the effect of labeled data scarcity indirectly. *Transfer learning* focuses on retaining knowledge gained while solving a problem from one domain and applying it to a different domain. Several research studies (3; 39) show that fine-tuning DCCN models originally trained on a large and diverse dataset of natural images using small datasets of medical images helped improve accuracy and decreased time for convergence of model training or fine-tuning. However, since real clinical images are quite heterogeneous not only in the diseased regions or lesions, but also in healthy or normal tissues patterns or background, these two approaches can only partially reduce the impact of small image datasets because they cannot increase the diversity of medical image cases in the data. One recent study (13) shows that the synthetic images generated by a generative adversarial network (GAN) were not recognized as synthetic by radiologists and that using these GAN-generated synthetic images helped increase the performance of a DCNN model in classifying liver lesions. Thus, using GANs is a promising solution to generate more synthetic image data across many application domains including medical imaging.

However, GANs are very complex machine-learning models; how they can efficiently generate large numbers of diverse or heterogeneous medical images, particularly depicting lesions or diseases against a wide variety of normal tissue backgrounds, has not been well

investigated. In Chapter 6, a new and more computationally efficient algorithm without GANs to generate synthetic image data on a diverse image base is discussed. The new algorithm extracts the annotated diseased region or lesion blobs, randomly redistributes these lesion blobs and projects them onto negative (healthy) images to optimally achieve seamless insertion. Unlike a GAN approach is limited by a small number of annotated positive lesions, our new method can include more diverse negative images that do not need expert annotations. In clinical practice the number of regular (disease-free) images are the majority and are easy to acquire for research purposes. Thus, this new algorithm can produce diverse synthetic image data generated using negative images with a broad diversity or heterogeneity of standard tissue patterns, which we hypothesize can help better train DCNN models and improve model performance. Therefore, the objective of this study is to test the feasibility of our hypothesis by comparing three deep transfer learning models (VGG-16 (40), ResNet-50 (17), and Inception V3 (43)) trained using a small set of original images and a large set of the synthetic images. Recently, these three deep learning models have been widely used in developing new CAD schemes of medical images with promising results.

1.2 Introduction to Hybrid Deep Learning

Medical image analysis has become increasingly important for accurate diagnosis, treatment, and disease management in modern healthcare. Deep learning models have proven to be highly effective in analyzing medical images. However, medical image analysis often involves complex tasks such as segmentation, classification, and generation of synthetic data, which require specialized deep learning models. To address these challenges, researchers have proposed hybrid and multistage deep learning models that combine different deep learning architectures with traditional machine learning techniques. These models leverage the

strengths of each approach to improve the accuracy and efficiency of medical image analysis tasks.

In medical image segmentation, hybrid deep learning models use a combination of deep convolutional neural networks (DCNNs), recurrent neural networks (RNNs), and unsupervised learning algorithms to segment specific regions of interest in medical images accurately. Multistage deep learning models use a two-step approach where a rough segmentation is first performed using a CNN, followed by refinement using a more accurate RNN. In medical image classification, hybrid models use a combination of CNNs and recurrent neural networks to classify medical images accurately. These models can handle multiple input modalities, such as images, text, and patient data, to provide a more comprehensive diagnosis. However these models tend to overfit on smaller datasets and complex for simple tasks like classification and segmentation

In medical synthetic data generation, hybrid models use a combination of GANs and CNNs to generate synthetic medical images that are highly similar to authentic medical images. These synthetic images can be used to train deep learning models, thereby increasing the size and diversity of the available dataset.

This dissertation is an overview of the state-of-the-art hybrid and multistage deep learning models for medical image segmentation, classification, and synthetic data generation and highlights their potential to revolutionize medical image analysis and improve the accuracy of diagnosis and treatment in healthcare.

Chapter 2

Related Work

This chapter explains the evolution of computer vision models, medical image analysis and hybrid model and the foundation on which this research is built.

2.1 Fundamentals of Deep Learning

Deep learning (DL) is a subfield of *machine Learning* (ML) that uses artificial neural networks with multiple layers to learn hierarchical representations of data. DL models have gained widespread popularity due to their ability to handle complex data and outperform traditional ML methods in various applications such as image recognition, natural language processing, and speech recognition. This chapter discusses some of the most commonly used DL models, their architectures, and their applications.

Recurrent neural networks (RNNs) (38) are another popular DL model that are commonly used in natural language processing and speech recognition. RNNs are designed to process sequential data and have a feedback mechanism that allows information to be passed from one time step to another. *Long short-term memory* (LSTM) and *Gated Recurrent Unit*

(GRU) are two types of RNNs designed to address the vanishing gradient problem that can occur in traditional RNNs. LSTMs use a more complex memory cell structure that allows them to store and retrieve information over longer periods of time. They have three gates: input, forget, and output gates. The input gate decides which information to let into the cell, the forget gate determines which information to discard from the cell, and the output gate decides which information to output. On the other hand, GRUs use a simpler architecture that combines the input and forget gates into a single update gate. They only have two gates: reset and update gates. The reset gate determines how to combine the new input with the previous memory, and the update gate decides how much of the previous memory to keep and how much of the new input to incorporate. LSTMs and GRUs have achieved state-of-the-art performance on various language modeling tasks, including machine translation, speech recognition, and text generation.

Transformers (46) are designed to process sequences of tokens, such as words or subwords, and are composed of multiple self-attention layers that allow the model to attend to different parts of the input sequence. The Transformers attention mechanism allows the model to capture long-range dependencies and has achieved state-of-the-art performance on various language modeling tasks, including machine translation, question answering, and summarization.

In summary, DL models have revolutionized artificial intelligence and achieved state-of-the-art performance on various applications, including computer vision, natural language processing, and speech recognition. CNNs, RNNs, GANs, and transformers are some of the most commonly used DL models, each with their unique architecture and applications. The rapid development of DL models and their applications is expected to continue, leading to significant advancements in AI in the years to come.

2.2 Fundamentals of Computer Vision

Computer vision (CV) is a subfield of artificial intelligence that focuses on enabling machines to interpret and understand visual data from the world around us. Computer vision models have made significant strides in recent years, and have been used in various applications such as object detection, image segmentation, and image recognition. This literature review will discuss some of the most commonly used CV models, their architectures, and their applications.

Convolutional neural networks (CNNs) (1) are a type of CV model that have been widely used for image classification tasks. CNNs are designed to process data with a grid-like structure, such as images, and are composed of multiple convolutional layers that extract features from the input data. These features are combined in fully connected layers to produce the final output. CNNs have achieved state-of-the-art performance on various image recognition tasks, including object detection and segmentation.

Region-based convolutional neural networks (R-CNNs) (14) are another type of CV model that have been widely used for object detection tasks. R-CNNs are designed first to generate region proposals, or potential object locations, using the selective search or similar methods. A CNN then processes these proposals to extract features, which are then fed into a classifier to determine the object's class in each proposal. *Mask R-CNNs* (16) are an extension of R-CNNs that can perform object detection and instance segmentation. In addition to generating region proposals and classifying objects, Mask R-CNNs can also output a binary mask for each detected object, indicating the precise boundaries of the object.

Generative adversarial networks (GANs) (15) are a type of computer vision model that can be used for image-generation tasks. GANs consist of two neural networks: a generator and a discriminator. The is trained to generate samples similar to the training data, while

the *discriminator* is trained to distinguish between real and fake samples. The generator and discriminator are trained simultaneously in a minimax game, where the generator tries to fool the discriminator, and the discriminator tries to identify the real samples correctly.

In summary, Computer Vision models have made significant strides in recent years, and have been used in various applications such as object detection, image segmentation, and image recognition. CNNs, R-CNNs, Mask R-CNNs, and GANs are some of the most commonly used computer vision models, each with their unique architecture and applications. The rapid development of computer vision models and their applications is expected to continue, leading to significant advancements in AI in the years to come.

2.3 Medical Image Analysis for Disease Diagnosis

Medical image analysis is an essential task in healthcare that enables clinicians to make accurate diagnoses and decisions for patient care. With the advent of deep learning techniques, medical image analysis has advanced significantly in recent years.

One of the most significant research areas in medical image analysis is detecting and classifying cancerous lesions. Various studies have proposed deep learning-based models for the early detection of lung, breast, and skin cancer, among others. For instance, researchers have developed a deep learning-based image processing model to detect lung cancer (47). Another study used deep learning models to diagnose breast cancer using mammography images, achieving an accuracy of 90%.

Besides cancer, deep learning models have also been used to diagnose diseases such as Alzheimer's, multiple sclerosis, and diabetic retinopathy. For instance, researchers (41) have developed a deep learning-based algorithm for diagnosing Alzheimer's disease using structural magnetic resonance imaging (MRI) images. The proposed model achieved high accuracy in detecting Alzheimer's disease in its early stages.

In addition to detecting and classifying diseases, deep-learning models have been used to segment medical images to isolate specific structures and lesions. For instance, researchers have developed a deep-learning model for segmenting brain tumors from MRI images (29). The proposed model achieved high accuracy and outperformed other traditional segmentation methods.

2.3.1 Image Segmentation

Image segmentation is dividing an image into multiple segments, each corresponding to a different object or region within the image. This task is essential in computer vision and has numerous applications including object detection, medical imaging, and autonomous driving. This section discusses some of the most commonly used computer vision models for image segmentation.

One of the most widely used models for image segmentation is the U-Net architecture (37), first proposed in 2015. U-Net is a fully convolutional neural network (FCN) consisting of contracting and expanding paths. The contracting path comprises several convolutional and pooling layers that progressively reduce the spatial dimensions of the input image while increasing the number of feature channels. The expanding path comprises several transposed convolutional layers that upsample the feature maps and restore the spatial dimensions of the output segmentation map. U-Net has been shown to achieve state-of-the-art performance on various medical image segmentation tasks.

(28) introduced the first fully convolutional neural network (FCN) for semantic segmentation, which paved the way for developing many subsequent segmentation models. The FCN approach enables end-to-end learning for segmentation tasks using deconvolutional layers to upsample the feature maps to the input image size. (49) proposed a deconvolutional network for semantic segmentation, which uses unpooling and deconvolution layers to upsample the

feature maps. The proposed network achieved competitive results on the PASCAL VOC 2012 dataset. (4) introduced dilated convolutions, which allow the network to increase its receptive field without decreasing the spatial resolution of the feature maps. The proposed network achieved state-of-the-art performance on the PASCAL VOC 2012 dataset.

(6) proposed an encoder-decoder network with atrous separable convolutions for semantic segmentation. The proposed network achieved state-of-the-art performance on the Cityscapes dataset. (34) proposed an attention mechanism for the U-Net architecture, allowing the network to focus on relevant input image regions during the segmentation process. The proposed network achieved state-of-the-art performance on the NIH pancreas segmentation challenge. (48) proposed a bilateral segmentation network (BiSeNet) for real-time semantic segmentation. The BiSeNet architecture consists of a spatial path and a context path, combined using a novel spatial attention module. The proposed network achieved state-of-the-art performance on several real-time segmentation benchmarks.

Another popular model for image segmentation is the DeepLab architecture (5), first proposed in 2016. DeepLab is also an FCN and is designed to produce dense pixel-wise predictions. DeepLab uses atrous convolutional layers, which allow the network to compute dense feature maps with a large receptive field, while preserving the spatial resolution of the input image. DeepLab also incorporates dilated convolutional layers, which further increase the receptive field of the network. DeepLab has achieved state-of-the-art performance on various image segmentation tasks, including semantic and instance segmentation.

2.3.2 Image Classification

Image classification is a fundamental task in computer vision that involves assigning a label or category to an image. Deep learning models have shown remarkable performance on image classification tasks in recent years. This chapter presents a literature review of some of the

most influential papers in this field. (25) introduced the AlexNet architecture, which was the first deep convolutional neural network (CNN) to achieve state-of-the-art performance on the ImageNet large-scale visual recognition challenge (ILSVRC) dataset. AlexNet used several techniques such as data augmentation, dropout, and ReLU activations, which became standard practices in deep learning.

(40) paper introduced the VGG architecture, which used very deep CNNs with small filters. VGG achieved excellent performance on the ILSVRC dataset, and its architecture has been widely used as a starting point for other CNNs. The Inception architecture (43), combined different filter sizes and pooling strategies to extract features at different scales. The Inception architecture was designed to be more computationally efficient than other CNNs and achieved excellent performance on the ILSVRC dataset. The ResNet architecture (17), which used residual connections to enable the training of very deep CNNs. ResNet achieved state-of-the-art performance on the ILSVRC dataset and has since become a standard architecture for many image classification tasks.

(19) introduced the DenseNet architecture, which used densely connected blocks to facilitate feature reuse and improve gradient flow. DenseNet achieved state-of-the-art performance on the ILSVRC dataset and is highly efficient in terms of both computational resources and memory usage. (44) proposed a new scaling method for CNNs that balances the network depth, width, and resolution to improve efficiency and accuracy. The resulting EfficientNet architecture achieved state-of-the-art performance on the ILSVRC dataset with significantly fewer parameters than previous models.

2.4 Hybrid Deep Learning

Hybrid deep learning models have emerged as a promising approach to improve the performance of traditional deep learning models. These models combine different types of neural

networks, such as CNNs, RNNs, and GANs, to leverage the strengths of each network and overcome their limitations. In recent years, numerous studies have been conducted on various hybrid deep-learning models. For example, some researchers have proposed CNN-RNN hybrid models for image and speech recognition tasks (42). These models use CNNs to extract features from images and RNNs to process sequential data, such as speech signals.

(18) proposed a hybrid deep learning model that combines a CNN and a LSTM network for predicting stock prices. The CNN extracts features from the stock price data, which are then fed into the LSTM for prediction. (11) proposed a hybrid deep learning model that combines a CNN and a RNN for sentiment analysis. The CNN extracts features from the text data, which are then fed into the RNN for classification. (45) proposed a hybrid deep learning model that combines a CNN and a LSTM network for customer churn prediction in the telecommunications industry. The CNN extracts features from the customer data, which are then fed into the LSTM for prediction.

(7) proposed a hybrid deep learning model that combines a CNN and a LSTM network for time series forecasting. The CNN extracts feature from the time series data, which are then fed into the LSTM for prediction. Additionally, the authors proposed a novel loss function that combines mean squared error and mean absolute percentage error to improve the model's performance.

2.4.1 Hybrid Convolution Neural Network

Hybrid CNN models have emerged as a promising approach for computer vision tasks, combining the strengths of different CNN architectures. One of the earliest hybrid CNN models, called the AlexNet model (26). This model combined traditional convolutional layers with fully connected layers and significantly improved image classification accuracy on the ImageNet dataset. AlexNet became the basis for later hybrid models like VGGNet, GoogLeNet,

and ResNet.

VGGNet (40), proposed by Simonyan and Zisserman in 2014, expanded upon AlexNet's approach by using smaller convolutional filters and deeper layers. This hybrid model showed improved accuracy on the ImageNet dataset, with the tradeoff of increased computational complexity. GoogLeNet (43), introduced the concept of *inception* modules, which combined multiple convolutional filters of different sizes in parallel. This approach significantly reduced the parameters and computational complexity while maintaining high accuracy. ResNet (17), introduced the concept of residual blocks, which allowed the model to learn residual connections between layers. This approach enabled the training of much deeper networks while avoiding the vanishing gradient problem.

2.4.2 Hybrid CNN models in Medical Imaging

Hybrid CNN models have gained attention in medical image analysis due to their ability to combine multiple CNN architectures with improving accuracy and reducing computational complexity. This section explores some of the research conducted on hybrid CNN models in medical image analysis.

One of the earliest applications of hybrid CNN models in medical image analysis was in the segmentation of brain tumors. (22) proposed a hybrid model that combined 3D U-Net and V-Net architectures, showing promising results in segmenting brain tumors on the BraTS dataset. Another application of hybrid CNN models is in the classification of skin lesions. (50) proposed a hybrid model that combined multiple CNN architectures, including VGGNet and ResNet, showing improved accuracy in the classification of skin lesions on the ISIC dataset.

Hybrid CNN models have also been applied to detecting lung nodules in CT scans. (2) proposed a hybrid model that combined multiple CNN architectures, including Inception

and ResNet, showing improved sensitivity and specificity in the detection of lung nodules on the LIDC-IDRI dataset.

In addition to combining multiple CNN architectures, hybrid CNN models in medical image analysis have also been combined with other deep learning models. For example, Roy et al. proposed a hybrid model that combined CNNs with *Variational autoencoders* (VAEs) (23) to generate synthetic medical images, showing promising results in the augmentation of medical datasets. Recent research in hybrid CNN models has focused on combining multiple architectures for better performance. In conclusion, hybrid CNN models have shown significant progress in computer vision tasks, and combining multiple architectures and deep learning models can lead to improved accuracy and reduced computational complexity.

However, despite the significant progress made in medical image analysis using deep learning models, several challenges remain. One of the main challenges is the requirement for large amounts of labeled data for training deep learning models. Obtaining such datasets can be time-consuming and costly, especially for rare diseases. Additionally, the black-box nature of deep learning models can make it challenging to interpret their results, which can be crucial in medical decision-making. Also, hybrid models are complex and need high computational resources to implement. Also, the lack of larger datasets in the medical field due to the fragmentation of the market and HIPAA regulations, extracting more information from smaller datasets is necessary. This way simpler hybrid models can yield better performance over a smaller quantity of data.

Chapter 3

Research Challenges

This chapter introduces some of the challenges of medical image analysis. Below is the list of research issues for disease diagnosis and prognosis predictions using medical image data.

3.1 Current state & challenges

This chapter discusses existing practices, limited research possibilities and various challenges in medical image analysis.

3.1.1 Lack of High-quality Data

Medical image data presents unique challenges for machine learning researchers and practitioners, including the limited availability of high-quality labeled data, large variability across imaging protocols and patient populations, class imbalance, ethical and privacy concerns, disease variability, and the need for specialized expertise. Overcoming these challenges requires collaboration and innovation across multiple fields, including computer science, medicine, and ethics.

- Limited availability of high-quality medical image datasets with ground truth annotations:

  Machine learning algorithms require large amounts of high-quality labeled data to train and validate models effectively. However, obtaining large amounts of labeled medical image data is challenging. Medical image datasets are often small, expensive to obtain, and require expertise to annotate. Additionally, the acquisition of labeled medical image data may require collaboration between multiple institutions, which can be time-consuming and logistically challenging.

- Large variations in imaging protocols, equipment, and patient populations across different healthcare institutions:

  Medical imaging protocols, equipment, and patient populations vary widely across different healthcare institutions, which can lead to inconsistent image quality and hinder the generalization of algorithms across sites. This variability can lead to challenges in algorithm development and deployment, as algorithms trained on data from one institution may not perform well on data from another institution.

- Imbalanced class distributions in medical image datasets:

  Medical image datasets often exhibit class imbalance, where certain classes may be overrepresented or underrepresented. For example, in a dataset of chest X-rays, the prevalence of pneumonia may be much lower than the prevalence of normal images. This imbalance can impact the accuracy and generalization of machine learning models, as algorithms may learn to overemphasize the majority class and underemphasize the minority class.

- Ethical and privacy concerns surrounding the use of patient data for research purposes:

Medical image datasets often contain sensitive patient information, which can make it difficult to obtain and share data. Additionally, ethical concerns surrounding patient privacy and confidentiality may limit access to medical image datasets. Researchers must adhere to strict data protection guidelines when handling medical image data, which can add complexity and delay the research process.

- Variability in disease manifestations and image appearances:

  Medical conditions can manifest differently across different patients, and this variability can make it challenging to accurately identify and classify certain medical conditions. For example, in a dataset of brain MRI scans, the appearance of a brain tumor may vary widely depending on the location and size of the tumor. This variability can make it difficult for algorithms to learn to accurately detect and classify medical conditions.

- The need for specialized knowledge and expertise in medical imaging:

  Interpreting and analyzing medical images requires specialized knowledge and expertise in anatomy, pathology, and radiology. Machine learning researchers and practitioners must work closely with medical experts to ensure that algorithms are accurately and appropriately trained and validated. This collaboration can add complexity and cost to the research process.

3.1.2 Lack of hybrid models

The lack of complex deep learning models in medical image analysis can result in reduced accuracy, limited features, limited generalization, increased annotation time, reduced interpretability, ethical and regulatory concerns, and infrastructure and resource requirements. Deep learning models, such as CNNs, can overcome these issues and provide accurate and reliable results for medical image analysis tasks. However, the deployment of deep learn-

ing models in clinical settings requires careful consideration of the ethical, regulatory, and resource requirements of the analysis process.

- Reduced Accuracy: Medical image analysis is a critical task that requires high accuracy and reliability. Traditional machine learning algorithms may not be able to effectively capture the complexity of medical images, which can lead to reduced accuracy and increased false positives/negatives. Deep learning models, such as CNNs, can capture complex features in medical images and improve the accuracy of the analysis.

- Limited Features: Medical images can contain a large amount of information including texture, shape, and intensity. Traditional machine learning algorithms may not be able to capture all the relevant features in the data, leading to reduced performance. Deep learning models can learn features automatically, reducing the need for manual feature engineering and improving the quality of the features extracted from the data.

- Limited Generalization: Medical image analysis tasks often involve data from different imaging protocols, equipment, and patient populations. Traditional machine learning algorithms may not be able to generalize well to new and unseen data, leading to reduced performance in real-world scenarios. Deep learning models can generalize well to new and unseen data, making them suitable for deployment in diverse clinical settings.

- Increased Annotation Time: Traditional machine learning algorithms often require manual feature engineering and annotation, which can be time-consuming and require domain-specific knowledge. Deep learning models can learn features automatically, reducing the need for manual annotation and speeding up the analysis process.

- Increased Interpretability: Deep learning models can provide insights into the most relevant features for a given task, allowing for increased interpretability of the results.

Traditional machine learning algorithms may not provide as much interpretability, making it difficult for medical professionals to understand the analysis results.

- Ethical and Regulatory Concerns: Deep learning models require large amounts of data to train and validate effectively. However, medical data is often sensitive and subject to strict data protection regulations. Researchers must maintain data privacy and confidentiality throughout the analysis process, which can be challenging.

- Infrastructure and Resource Requirements: Deep learning models require significant computational resources, including powerful GPUs and large amounts of memory. Additionally, the training and deployment of deep learning models require software engineering and data science expertise, which can be challenging to acquire in the medical domain.

In summary, while deep learning models have shown great potential in medical image analysis, their transferability to other domains can be challenging due to issues related to data availability, domain shift, ethical and legal concerns, variability in clinical practice, and clinical relevance. Addressing these challenges requires collaboration between researchers, clinicians, and regulatory bodies to ensure that the models are safe, effective, and clinically relevant.

3.2 Research Objectives

Since the above-mentioned challenges are complex and need collaboration between research domains, this dissertation focuses on finding alternative solutions to achieve high quality results. Chapter **??** focuses on creating a hybrid deep learning model that simultaneously performs image segmentation and classification even with smaller amounts of data compared

to its individual component models. Chapter **??** proposes a two-stage deep learning architecture for image classification using transfer learning. This architecture classifies the correct annotated images with high accuracy. Chapter **??** focuse on incorporating image processing techniques to generate synthetic images from smaller original dataset. This method helps in creating larger and more diverse dataset for deep learning architectures.

Chapter 4

Novel Joint Deep Learning Model to Improve Image Segmentation and Classification

This chapter discusses a novel Joint-Net proposed to perform image segmentation and classification simultaneously using a skin cancer image dataset. This study is published in SPIE Medical Imaging, 2023 under the section of Computer Aided Diagnosis. My contributions include domain research, architecture design, literature review, experimentation, and writing.

4.1 Proposed Method

4.1.1 J-Net (Joint-Net)

A novel architecture, J-Net (33), is developed by combining the U-Net architecture with a binary image classifier, inspired by Y-Net (31). The J-Net architecture contains two-way feature learning modules. The first module is a U-Net (37), a deep down-sampling-to-up-sampling 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 37. 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 single sigmoid layers are connected to the classification convolution blocks. Thus, the feature maps from the encoder phase of the J-Net serves as the input to the binary classifier and the classification branch categorizes the image into two classes.

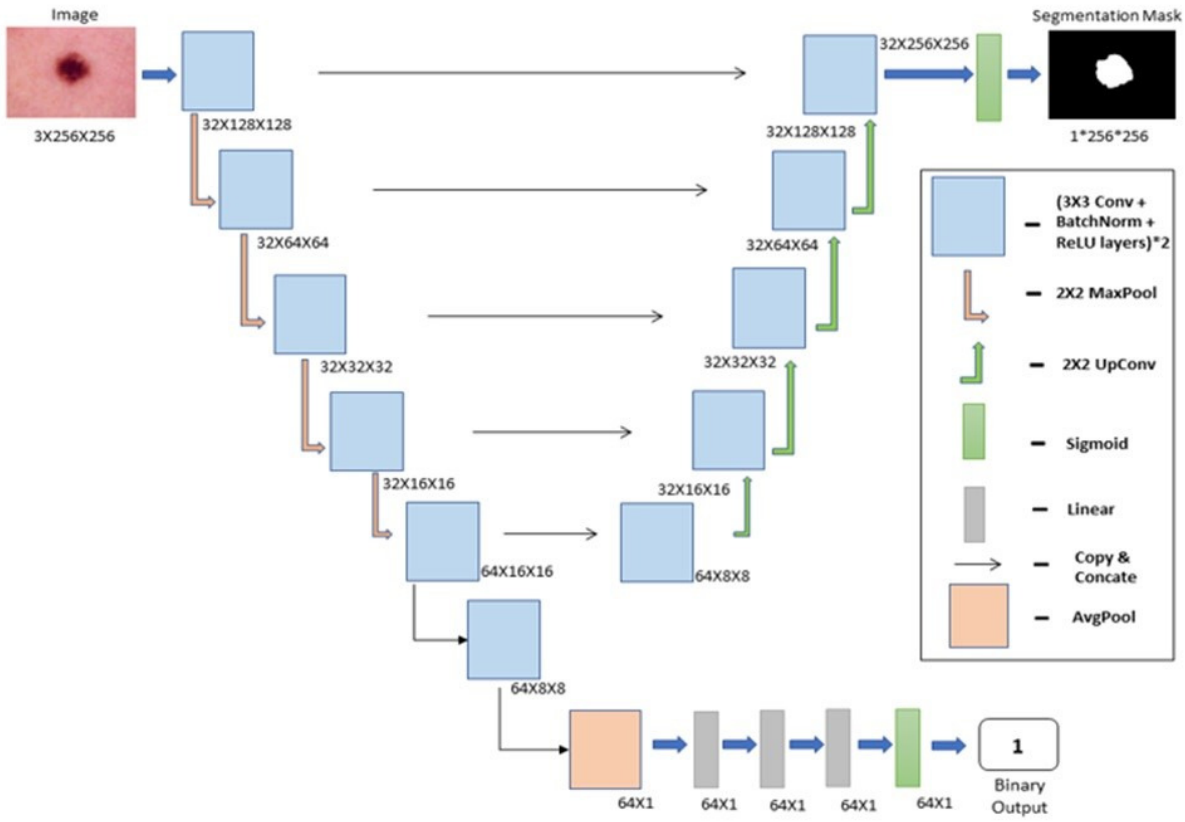

Figure 4.1: J-Net Architecture

The J-Net is optimized using the combined loss from segmentation (LossSeg) and classification tasks (LossClass). Combining losses helps the network understand each data label's features and its segmentation mask. The additional information learned by the classification branch helps the segmentation branch optimize the masks and vice versa.

$$Loss_{Total} = Loss_{Seg} + Loss_{Class} \tag{4.1}$$

4.1.2 Binary Image Classifier

A simple binary classifier is the second independent model used to perform image classification. The classifierâs 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 â1â meaning to severe and â0â meaning to the mild presence of melanoma on the lesions. For the model training, an image dataset of size 1200 with a 70:20:10 data spit for validation and test sets is drawn and the performance metrics are measured to compare with the results from the J-Net.

4.2 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 as shown in the 4.2. 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 âmildâ and 1 representing the âsevereâ presence of nevus cells in the corresponding image of a skin cancer lesion. In total, 1200 images are selected with equal distribution of both cases for the study.

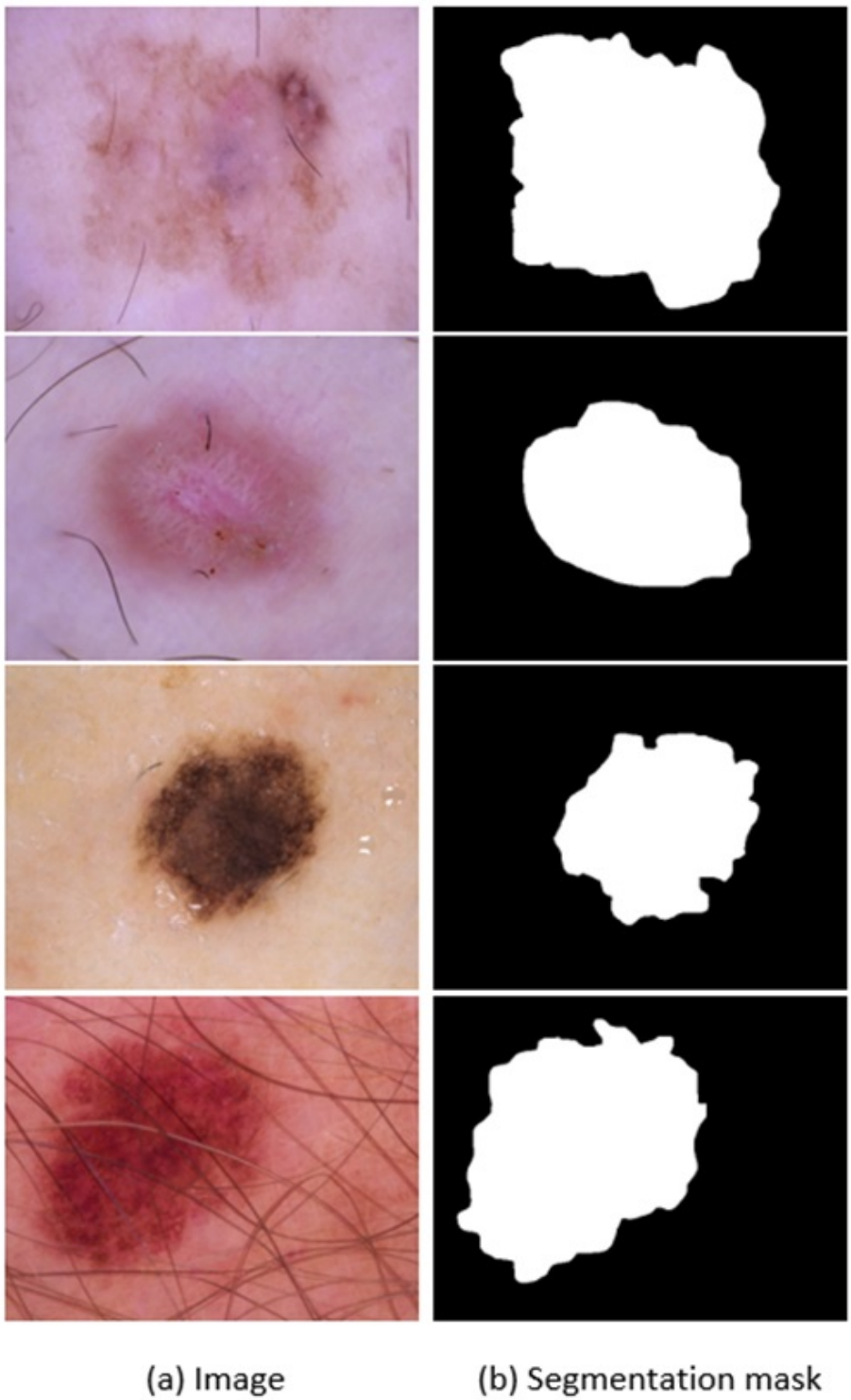

Figure 4.2: Skin Cancer Images and corresponding segmented lesions

4.3 Results

The results of the U-Net and the J-Net models are compared based on lesion segmentation and classification performance metrics. For testing purposes, 10 percent of the dataset is used. From the observed results 4.1, the J-Net model has higher segmentation accuracy than the U-Net model over smaller datasets. Especially, the J-Net model accurately segments the contours of the images. Both models segmentation metrics improved as the dataset size increased from 200 to 1200 images.

To compute the statistical significance, both the J-Net model and the U-Net models are evaluatued using the *Binary Cross Entropy* (BCE) loss. The BCE loss function measures the difference between the predicted binary segmentation and the ground truth binary segmentation. The models are trained 30 times and the test set BCE loss is observed. The mean distribution of BCE loss for J-Net is lower than for U-Net in the segmentation task showing over 8% improvement 4.4.

Similarly, the J-Net classifier outperformed the binary classifier with the same architecture with over 5% improvement when evaluated using the BCE loss 4.5. This experiment shows that adding additional information to training models can help them understand the hidden patterns from all perspectives.

The performance metrics for the lesion segmentation i,e., dice score, intersection over union (IOU), accuracy, and precision, are recorded between U-Net and J-Net. For the image classification task Recall and F1-score metrics are calculated to perform the comparison between a binary classifier and a J-Net classifier. (Refer to appendix 9)

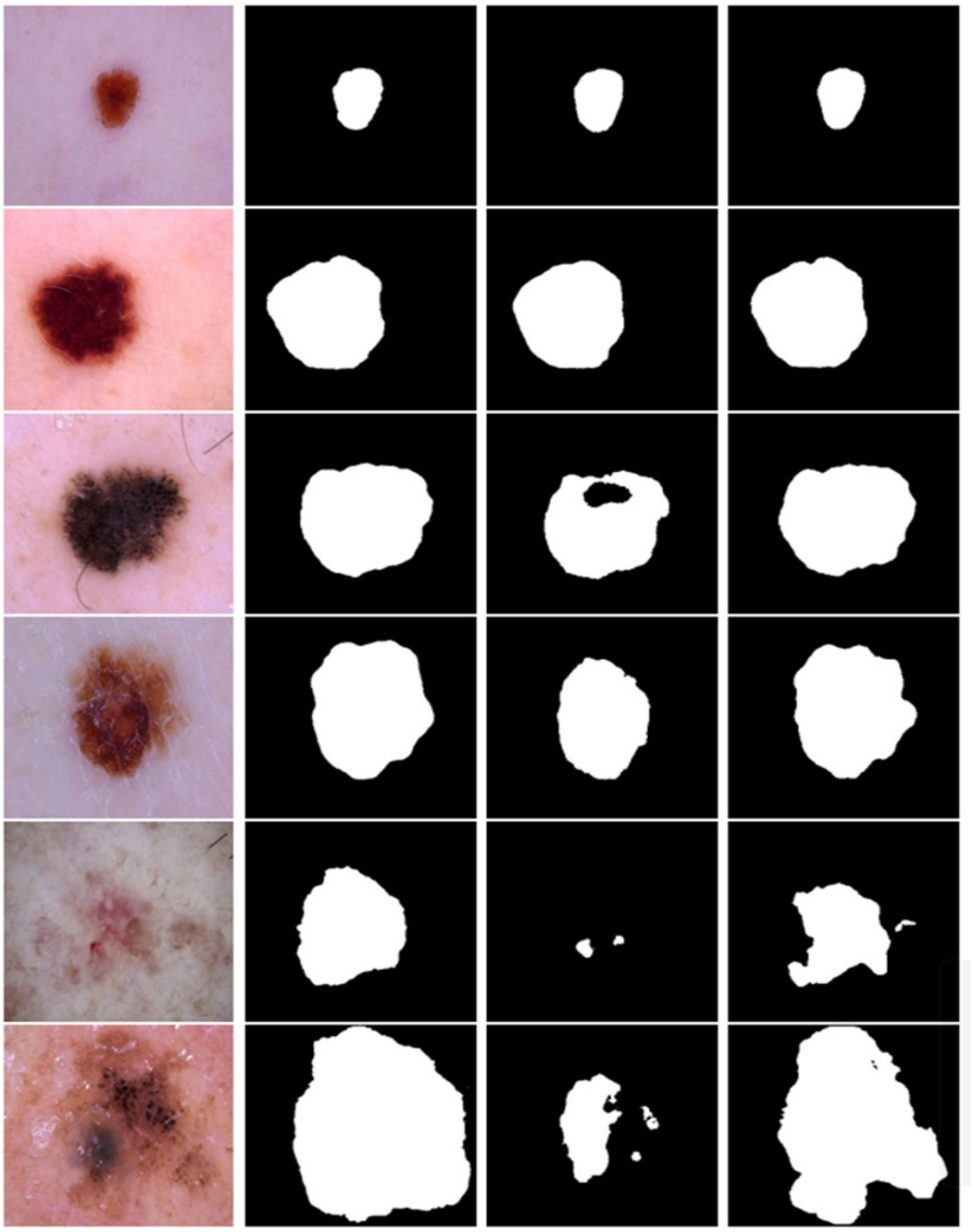

Figure 4.3: From left to right: skin cancer images, corresponding original segmentation masks, U-Net generated masks, and J-Net generated masks

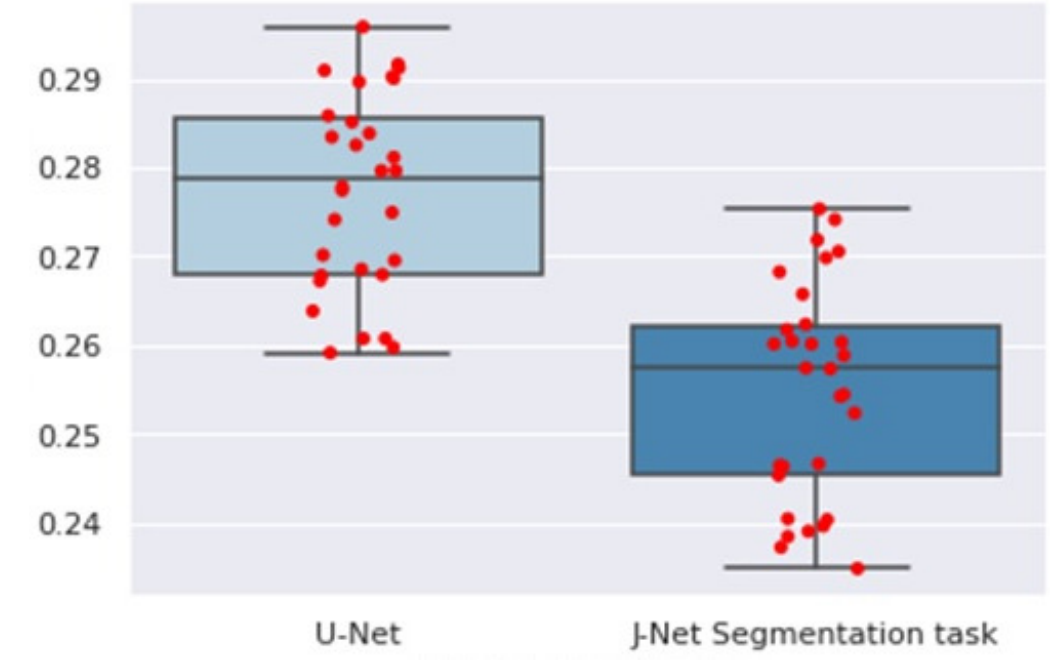

Figure 4.4: Boxplot results of BCE loss for U-Net vs J-Net Segmentation

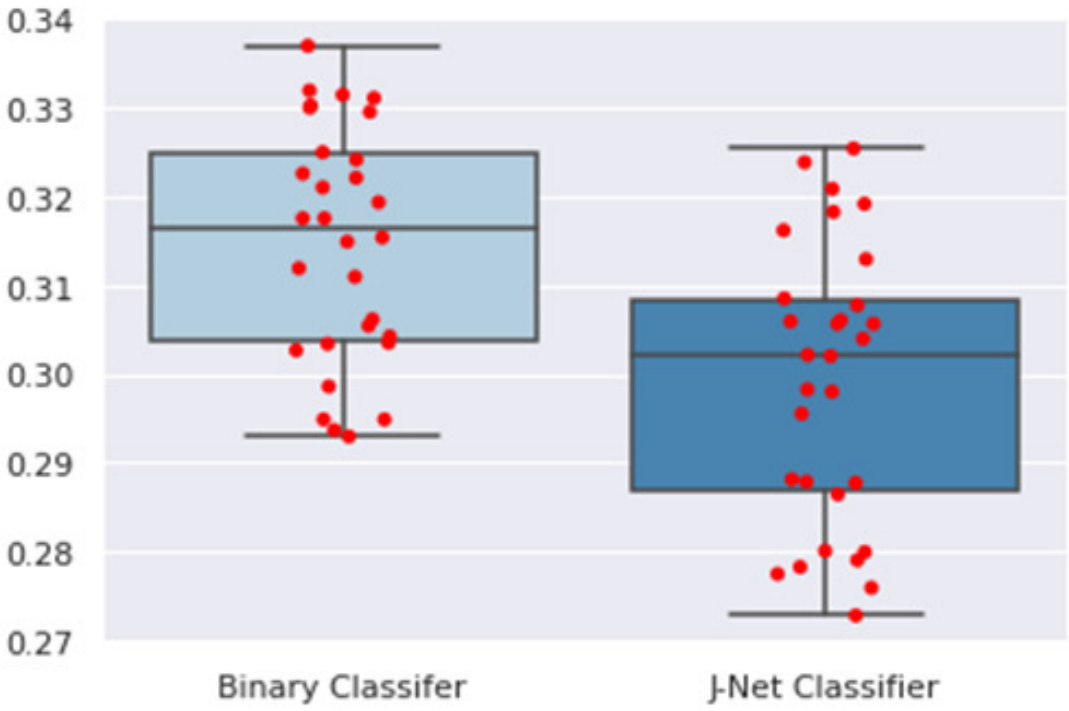

Figure 4.5: Boxplot results of BCE loss for Binary Classifier vs J-Net Classifier

	U-Net Metrics				J-Net Metrics			
Size	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

Table 4.1: Segmentation metric results of U-Net and J-Net with varying training data size from 200 to 1200

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 fairness. The outcomes of the J-Net classifier outperformed the binary classifier with better precision and recall, thus, it 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.

	Binary Classifier Metrics				J-Net Classifier Metrics			
Size	Precision	Accuracy	Recall	F1	Precision	Accuracy	Recall	F1
1200	0.74	0.90	0.73	0.73	0.78	0.93	0.77	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	0.68	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	0.64	0.69	0.82	0.68	0.67
200	0.66	0.80	0.65	0.64	0.68	0.82	0.66	0.66

Table 4.2: Classification metric results of Binary classifier and J-Net classifier with varying training data size from 200 to 1200

4.4 Discussion

In this chapter, 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:

- The J-Net joint model yields higher lesion segmentation and classification accuracy than using two single independent models, even with small datasets (as shown in 4.1 and 4.2). This indicates that different deep learning models may have different foci in learning and extracting image features, thus the generated segmentation results and/or classification scores are not highly correlated. As a result, this can also be explained due to the features learned from two integrated branches of the J-Net model, which enables improved model performance in both lesion segmentation and classification. Therefore, combining two tasks of lesion segmentation and classification into one model allows for a better understanding and integration of image features and the context learned from

different perspectives.

- This study also demonstrates a clear trend of performance increase of the deep learning models including U-Net and joint J-Net models as the increase of training and testing dataset size (i.e., from 200 to 1200 in this study). Observation of such an increasing trend clearly indicates that increasing training dataset size and diversity plays a very important role in developing more accurate and robust deep learning models using medical images. Thus, in future research, more effort should be added to increasing training dataset size by either collecting more clinical images or developing more effective algorithms or models to produce more clinically relevant synthetic images.

With several limitations including using only 2D images of skin cancer and difficulty to obtain accurate lesion segmentation masks, we recognize that this is a quite unique study to address two important issues in effectively applying deep learning models in medical image research. Thus, distinctive studies are needed to investigate more robust and muti-usage of data for different medical imaging applications in the future.

Chapter 5

A Novel Two-stage Deep-Learning Model to Improve Accuracy in Detecting Retinal Fundus Images

This chapter discusses a two-stage deep learning model to improve the detection accuracy using retinal fundus images. The objective of model is to automatically detect retinal fundus images without any artifacts. This study is published in SPIE Medical Imaging, 2022, San Diego. My contributions include domain research, algorithm design, literature review, and writing.

5.1 Proposed Method

Transfer learning is used to build new deep-learning models to classify the retinal fundus images into three classes. For this purpose, the ResNet-50 (17) architecture is selected to build our transfer learning model, which has been pre-trained using an extensive ImageNet

database. Although many Deep Convolutional Neural Network (DCNN) models have been developed and used as transfer learning models in medical imaging informatics, a DCNN model using ResNet-50 architecture is chosen for this study. ResNet-50 has several advantages compared with VGG19 (40) 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.

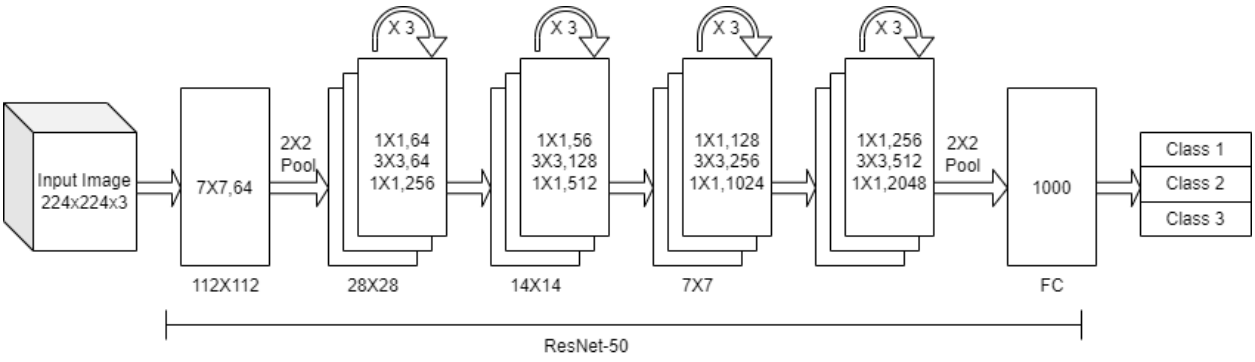

Figure 5.1: Block Diagram of Stage1 Transfer Learning with ResNet50

This study picks 50 images from each class to build a training dataset. Thus, 150 images are used to fine-tune the ResNet-50 model (17). 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 with the highest probability score.

However, the conventional classification method shown in the above figure based solely on the highest probability score for the three classes may not achieve adequate classification performance. Thus, to improve model performance, a new classification method is proposed shown in Figure 5.5 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 where the model-generated probability score exceeds 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 âdifficultâ images. In order to further classify these difficult images, a second transfer learning model is added. 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 âeasyâ 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.

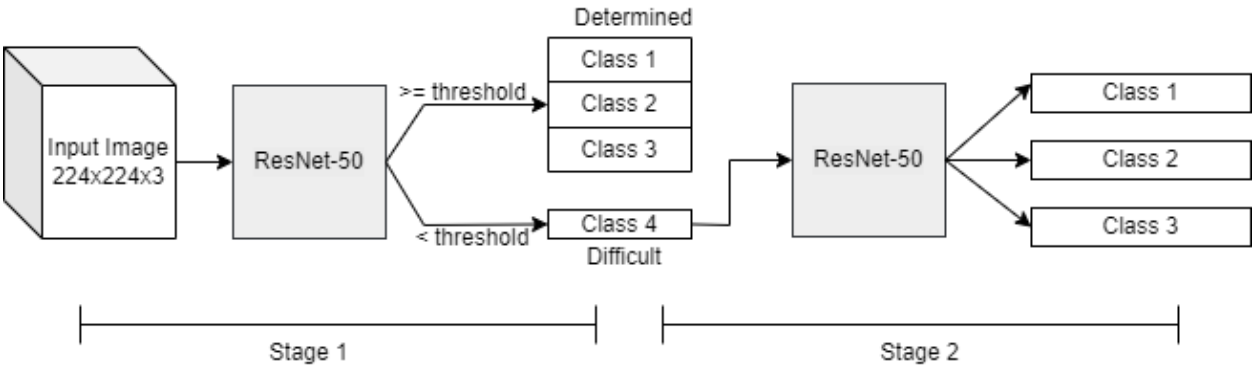

Figure 5.2: Block diagram of the proposed 2-stage model

5.2 Dataset

The Google search engine is queried and retrieved 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, as shown in Figure 5.1. 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 5.2. 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 5.3. Based on the above criteria, Class 1 has 620 images, Class 2 has 134 images, and Class 3 has 473 images.

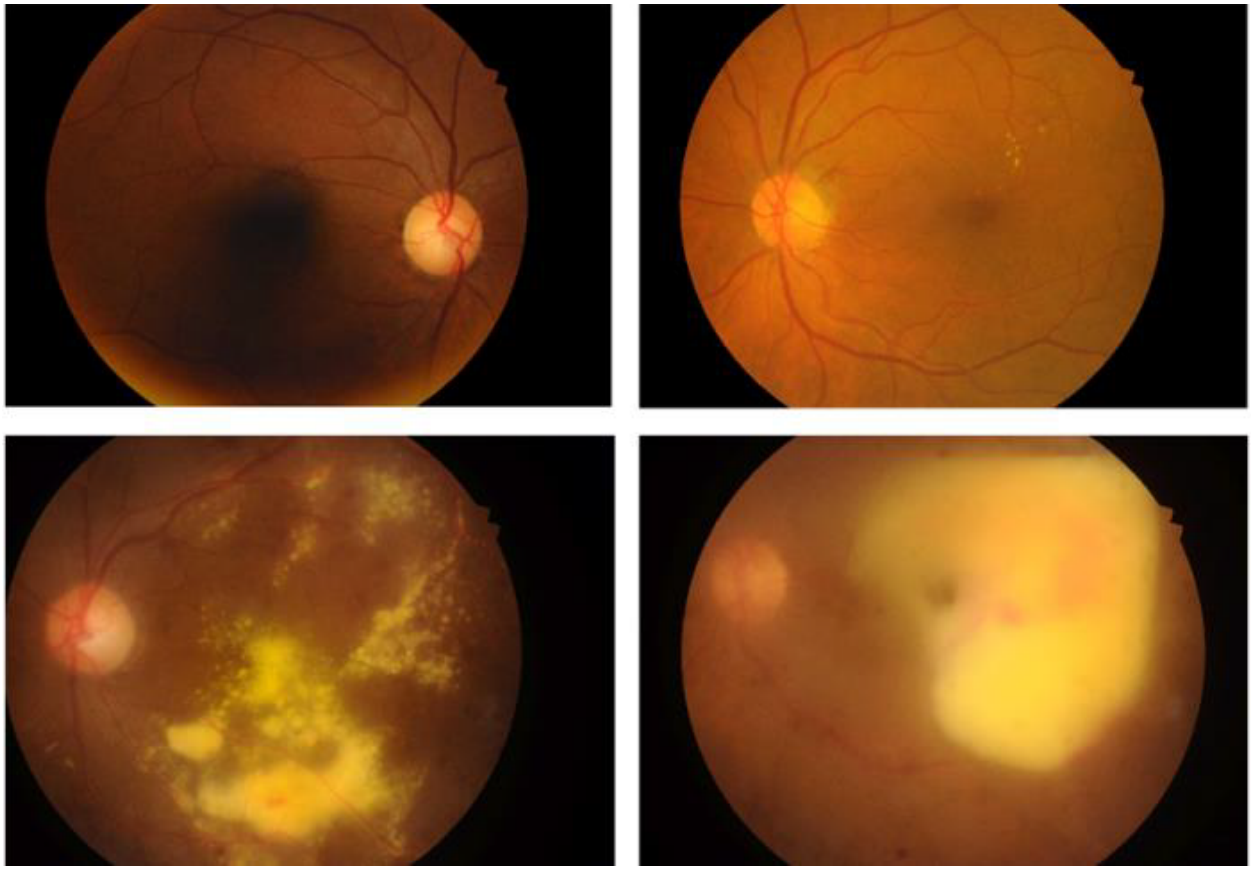

Figure 5.3: Example images for Class 1

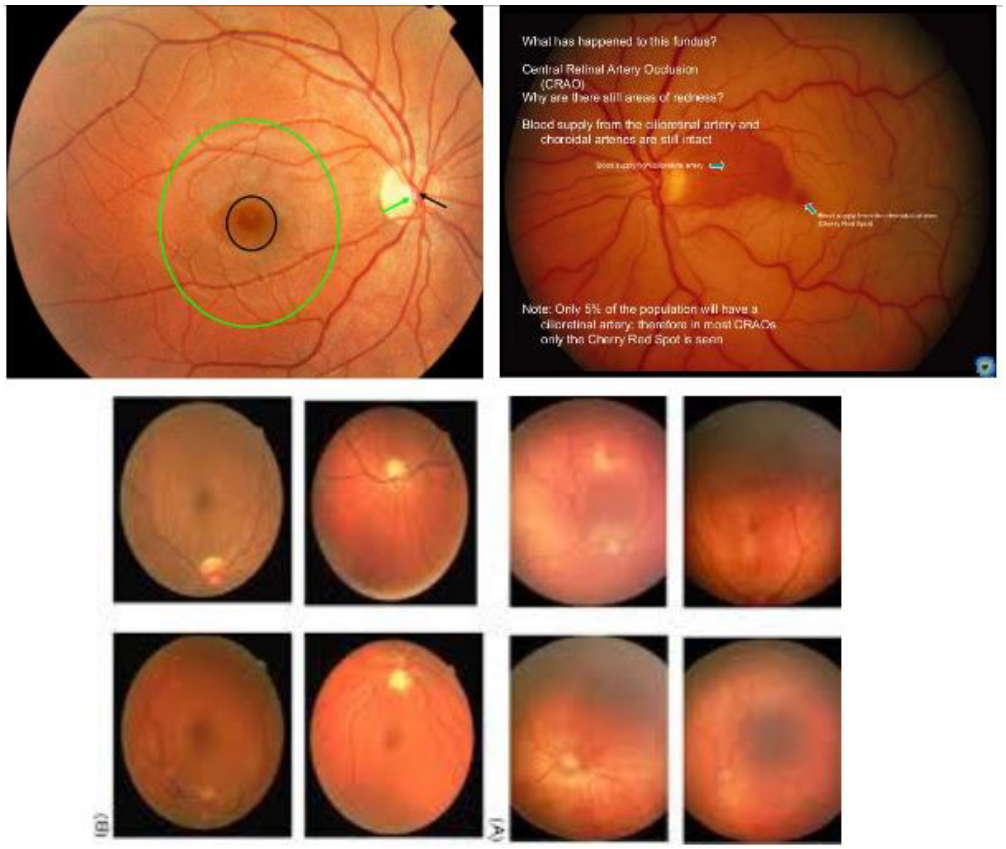

Figure 5.4: Example images for Class 2

Figure 5.5: Example images for Class 3

5.3 Results

Table 5.1 shows the performance when using one ResNet-50 model to classify images into three classes without adding a threshold value on the SoftMax layer. Overall 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 and positive predictive values are summarized in Table 5.2.

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

Table 5.1: Stage 1 results threshold on the SoftMax Layer

After adding threshold values on the SoftMax layer of the model, the easy images will be

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
Positive Predictive Value(PPV) of 3 Classes	0.933	0.945	0.961	0.970	0.966

Table 5.2: Performance of model in Stage 1 using different threshold values in SoftMax layer

classified into 3 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 decrease due to the increase of undetermined images in Class 1 (as shown in Figures 5.6 and 5.7).

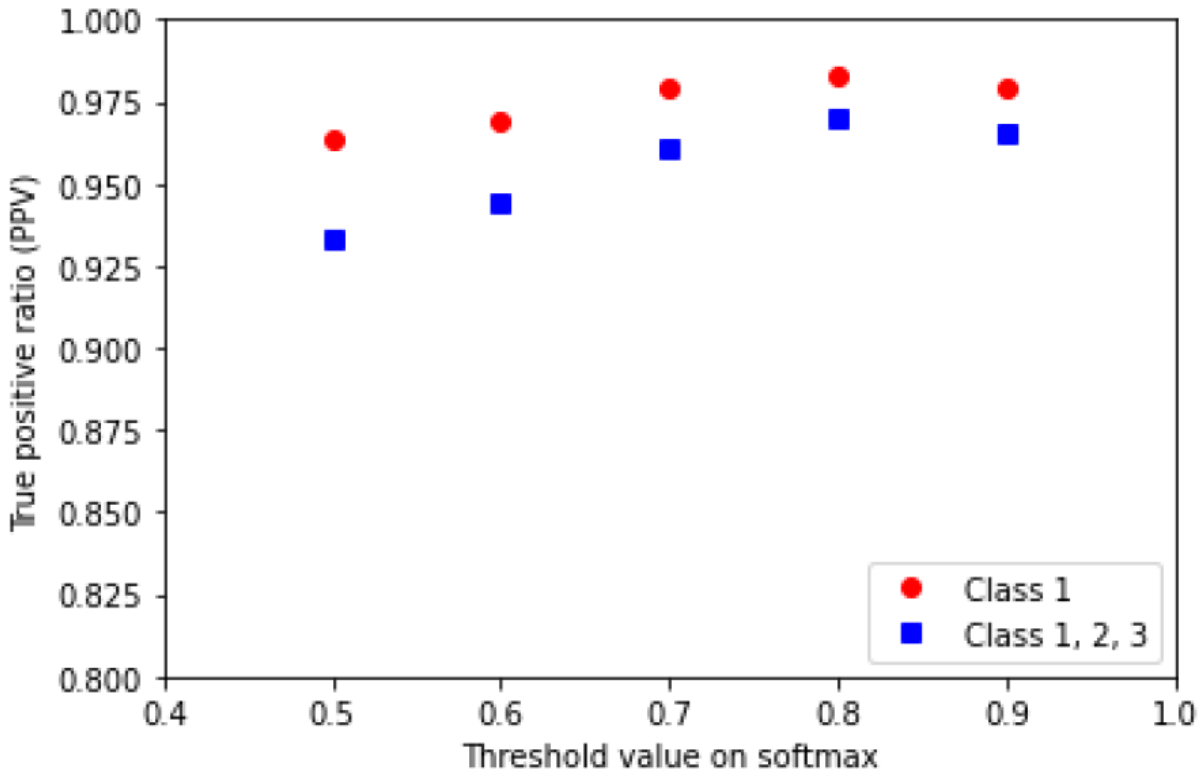

Figure 5.6: True positive ratio (PPV) vs Threshold value

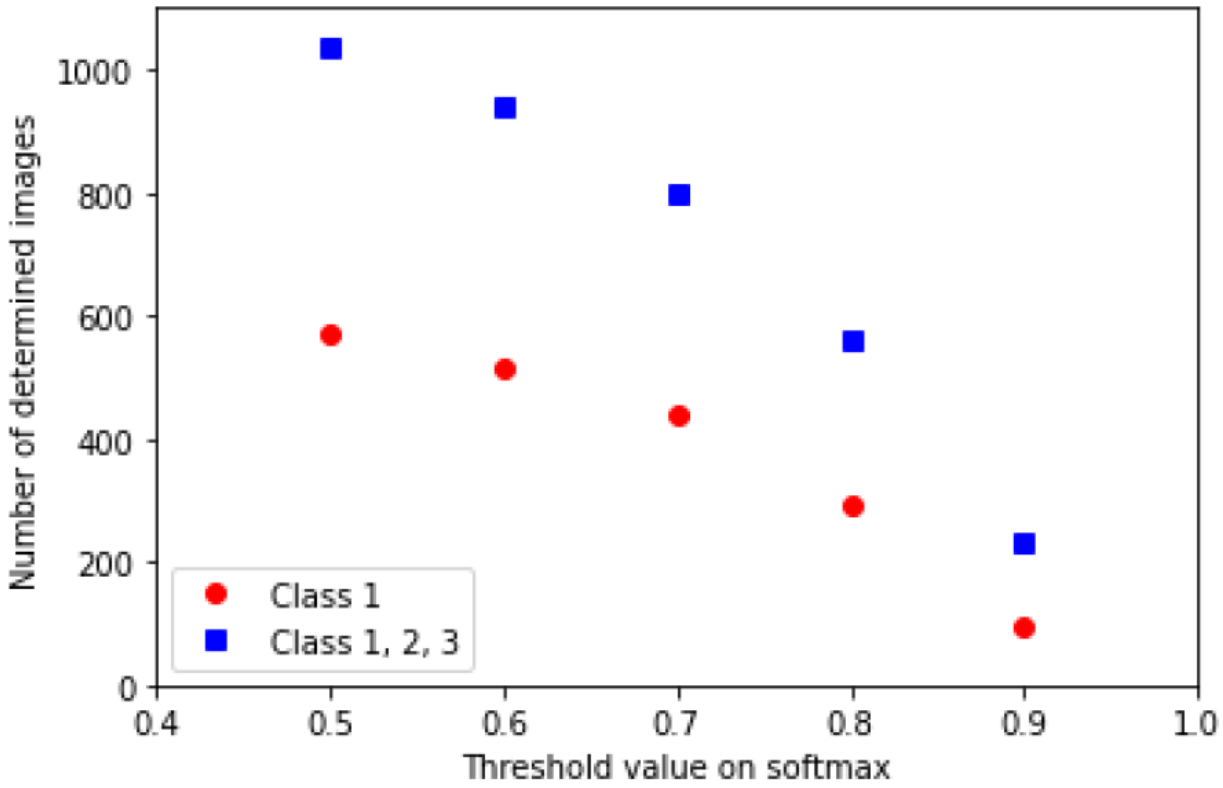

Figure 5.7: Number of determined images vs Threshold value

Table 5.3 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.

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
Positive Predictive Value(PPV) of 3 Classes	0.950	0.930	0.963	0.962	0.979

Table 5.3: Performance of model in Stage 2 using different threshold values in SoftMax layer

Table 5.4 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 5.4 and Table 5.1), the 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 two-stage model increases the number of true positives for all three classes from 990 to 1,029. However, the total number of true positives for Class 1 remains the same using the single-stage and two-stage models.

Class	1	2	3	Total
Assigned Images(Stage: 1 + 2)	571	77	417	1065
True Positive(TP) images	563	52	410	1029
False Positive(FP) images	8	25	7	36
Positive Predictive Value(PPV)	0.986	0.675	0.983	0.966

Table 5.4: 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.

5.4 Discussion

This research introduced, created, and tested a novel two-stage model to classify retinal fundus images, utilizing transfer learning ResNet-50 models. Our experimental results revealed several crucial findings:

Firstly, the two-stage transfer-learning model outperforms the traditional single-stage model, as evidenced by an improvement in the positive predictive value (PPV) from 91.9% to 96.6% (an increase of 4.7% compared to the base model) when considering the entire dataset's key statistical parameters (Tables 5.1 and 5.4).

Secondly, our study primarily focuses on identifying complete retinal fundus images (Class 1). Comparing the results of Class 1 between the single-stage and two-stage transfer-learning models, an evident decrease in false positives is observed, from 25 to 8, even though the total number of identified images is the same.

Thirdly, the first stage of our proposed model plays a vital role in identifying the optimal threshold to improve potential samples required for Stage 2. Increasing the threshold from 0.8 to 0.9 does not contribute to increasing the PPV value, as shown in Figure 5.6 and Table 5.2.

Furthermore, similar to conventional machine learning or computer-aided detection schemes, a deep-learning model's performance, heavily relies on the content distribution of training datasets. As such, identifying or classifying difficult or subtle images using a deep-learning model trained or fine-tuned using a small dataset that cannot sufficiently represent the difficult images or outliers can be challenging. Developing a two- or multi-stage deep-learning model or scheme offers significant advantages to address and solve this challenge, which is our primary contribution to the medical imaging informatics or CAD field.

The future expansion is to explore this phenomenon of integrating image-processing techniques with deep-learning frameworks using a more diverse dataset of retinal images and investigate the application of integrated architectures that combine traditional image processing and deep-learning frameworks.

Chapter 6

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

This chapter discusses an efficient synthetic data generation model to improve image analysis. The objective is to generate more data samples with the existing annotated images. The patterns from the originals images are used to synthesis new data. My contributions include domain research, model design, literature review, and writing.

6.1 Dataset

This study is performed on the Indian Diabetic Retinopathy Image Dataset (IDRiD), a publicly available dataset of retinal fundus images (36). The dataset consists of retinal

images of 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, clinicians have annotated all three types of diseased lesions in the images. This dataset is randomly divided into two independent training and testing 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 3 or 2 annotated mask images, one for each type of disease in Categories 1 and 2, respectively. Figure 6.1 shows an example image in Category 1 and three corresponding masks of HE, EX, and SE lesions. Figure 6.2 shows another example image of Category 2 and the two associated masks of HE and EX lesions. All labeled mask images are pseudo binary images without lesion density information.

The following image preprocessing steps are performed to clean the data. 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 6.3 shows an example of this image preprocessing result. 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, extra padding is added to generate another set of images with square ROI of size 299×299 pixels, which is the input image size for

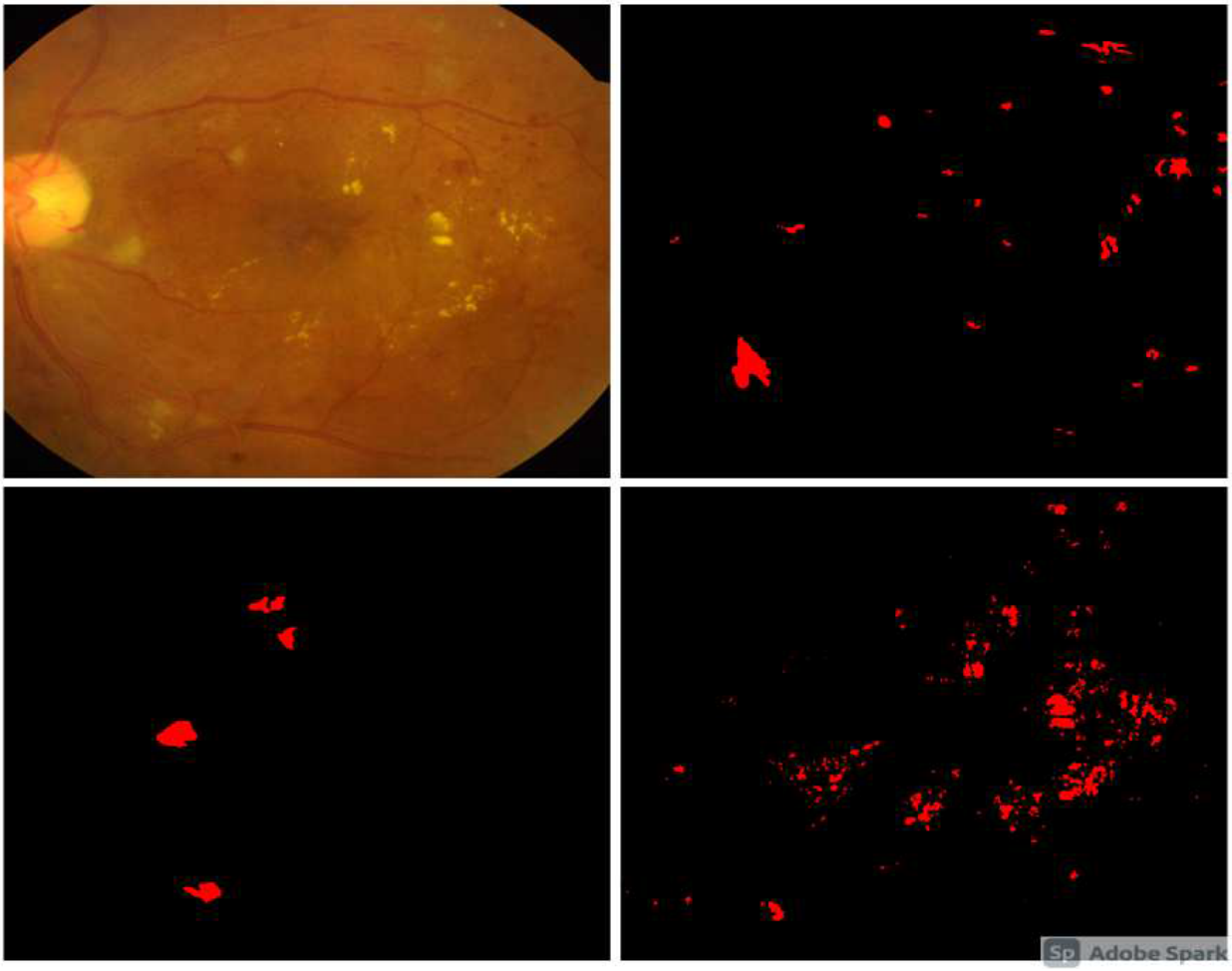

Figure 6.1: Original retinal image (top left) with three annotated masks for three types of lesions (top right-clockwise: HE, EX, and SE).

Inception v3.

6.2 Sythetic Data Generation Method

To generate synthetic images for this study (24), another dataset of 60 healthy (or normal) retinal fundus images, which an ophthalmologist in our local clinic provided as a base to generate images with synthetically added diseased patterns or lesions are used. Then the following new synthetic data generation algorithm is designed and applied to extract lesion

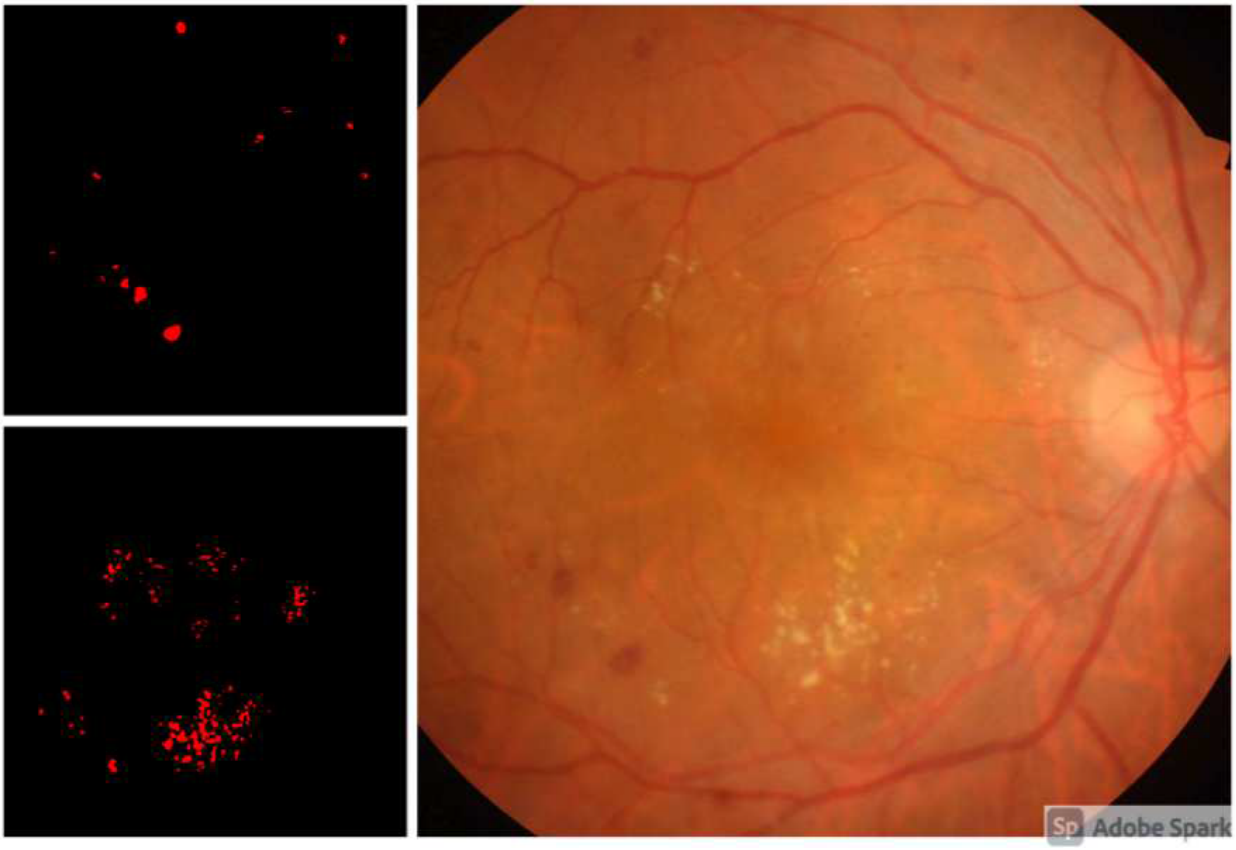

Figure 6.2: Original retinal image (right side) with two annotated masks for two types of lesions (top left: HE; bottom left: EX).

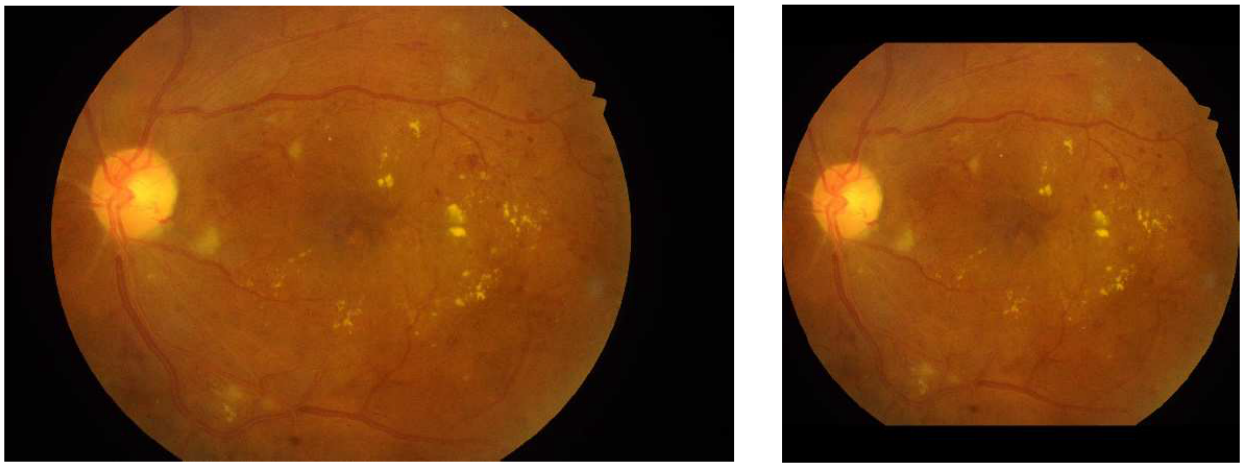

Figure 6.3: Original retinal image (left side); image after pre-processing (right side).

blobs, randomize their distribution to create new lesions, and project the synthetic lesions onto these healthy retinal fundus images. The algorithm includes the following six major steps.

1. Because there are three separate disease masks of HE, EX, and SE lesions in Category 1 images (as shown in Figure 6.1) and two separate disease masks of HE and EX lesions in Category 2 images (as shown in Figure 6.2) in the IDRiD dataset, a combined mask image is created that contains either three types of lesions for the Category 1 disease case or two types of 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 6.4 shows one combined mask that contains 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, the number of lesion blobs (N) is counted in each mask image and label them from 1 to N. After examining the whole dataset, it is noticed that the 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, a random selection of 2 to N blobs from each originally combined mask to generate 20 new masks is made. Thus, in this step, a total of 1,080 new mask images (54 combined masks $\times$ 20 new masks per combined mask) are generated that each containing a random number of lesion blobs. Figure 6.4 shows an example of one originally combined lesion mask and 20 new lesion masks generated from this combined lesion mask.

3. An element-wise multiplication is performed 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, the pseudo blobs in the masks generated in Step 2 back to the original retinal fundus images to extract real lesion blobs are projected. As a result, the new mask images contain real blobs that contain lesion density information.

4. Each real lesion blob mask obtained in Step 3 is flipped horizontally and vertically. Thus, one real lesion mask becomes three (original, horizontally flipped, and vertically flipped) with the same set of real lesion blobs in different positions and orientations. As a result, 3,240 real lesion masks (1,080 generated mask images $\times$ 3 transformation options) are generated. Due to the 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. To overcome this data imbalance issue, a multiple angular rotations are performed to increase the numbers of masks in the five categories to generate approximately comparable numbers. Table 6.1 shows the resulting number of the final synthetic lesion blob masks in each of the seven categories. While we did not need to generate datasets for all seven categories for the experiments presented herein, we did so to demonstrate the algorithm's ability to generate data in categories beyond those that appear in the original dataset.

5. Finally, each lesion blob mask is inserted and projected from Step 4 onto a randomly chosen healthy retinal fundus image. The projection takes the overlapping pixels between the healthy retinal fundus image and the synthetic lesion blob mask from the

synthetic lesion blob mask. In contrast, the remaining pixels are extracted from the healthy fundus image. As a result, 7,092 synthetic images of Category 1 and 6,786 synthetic images of Category 2 are generated, as shown in Table 6.1, which were then used to train deep learning model as described below.

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

Table 6.1: Numbers of synthetic real lesion blob masks generated in seven categories of lesion distributions

In summary, Figure 6.4 illustrates the step-by-step workflow of the proposed synthetic image data generation algorithm. The figure shows an original retinal fundus image (Figure 6.5a) and a corresponding lesion mask (Figure 6.5b). These are taken from the IDRiD dataset, a randomly generated lesion mask (Figure 6.5c), a lesion blob image that is obtained after projecting the mask with the corresponding original image (Figure 6.5d), a sample flip (horizontal, Figure 6.5e), a random angular orientation (Figure 6.5f), and a lesion blob mask insertion to a healthy retinal fundus image to generate the final synthetic image (Figure 6.5g). Figure 6.6 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.

6.3 Experiments

In this study, six deep-transfer-learning DCNN models with three different architectures are used, aiming to classify retinal fundus images into two different disease categories. Although

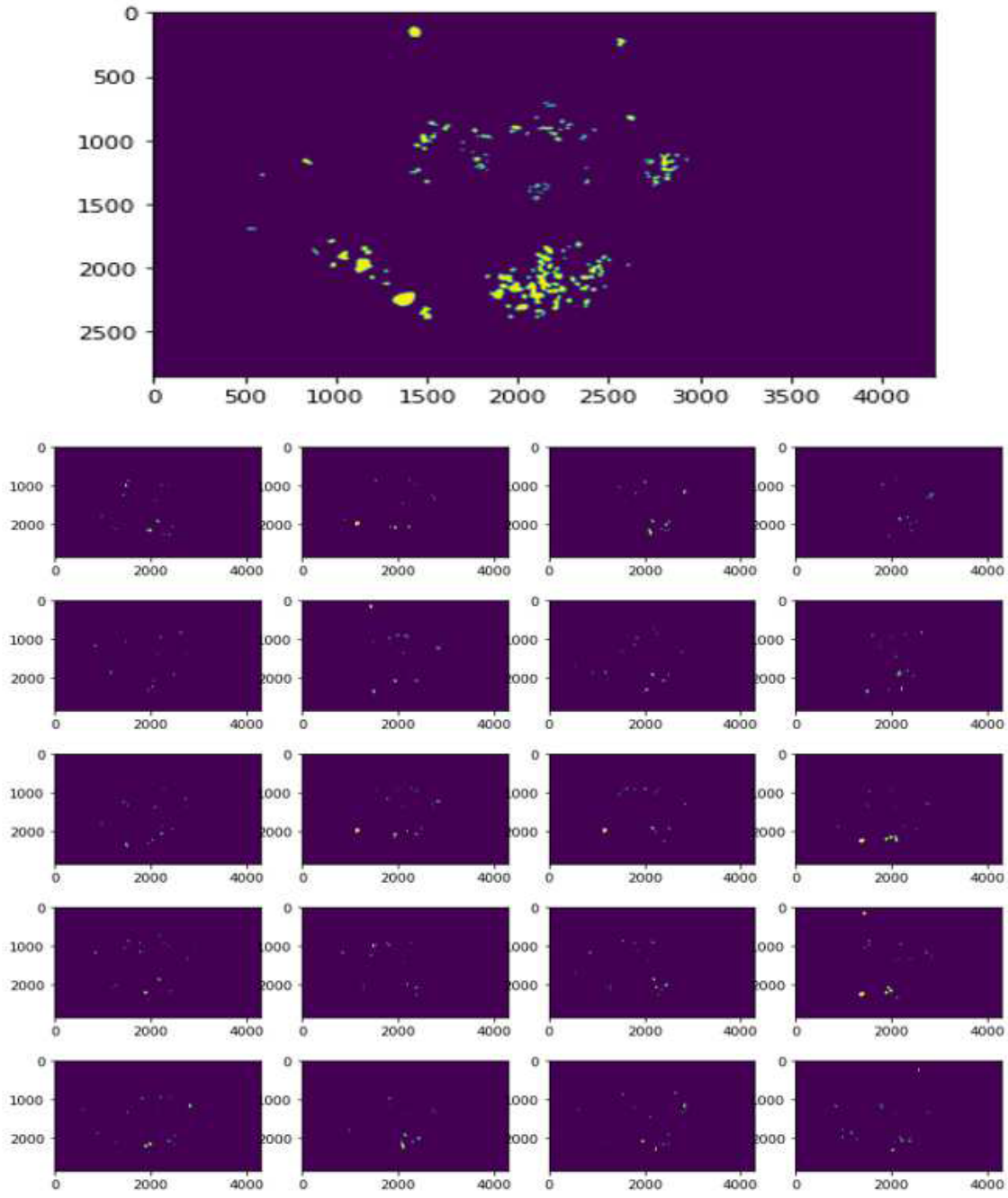

Figure 6.4: The image on top shows one originally combined mask image that consists of all 3 types of lesions (Category 1) and images below on the grid show 20 new mask images generated based on the top combined mask image with random distributions of lesion blobs.

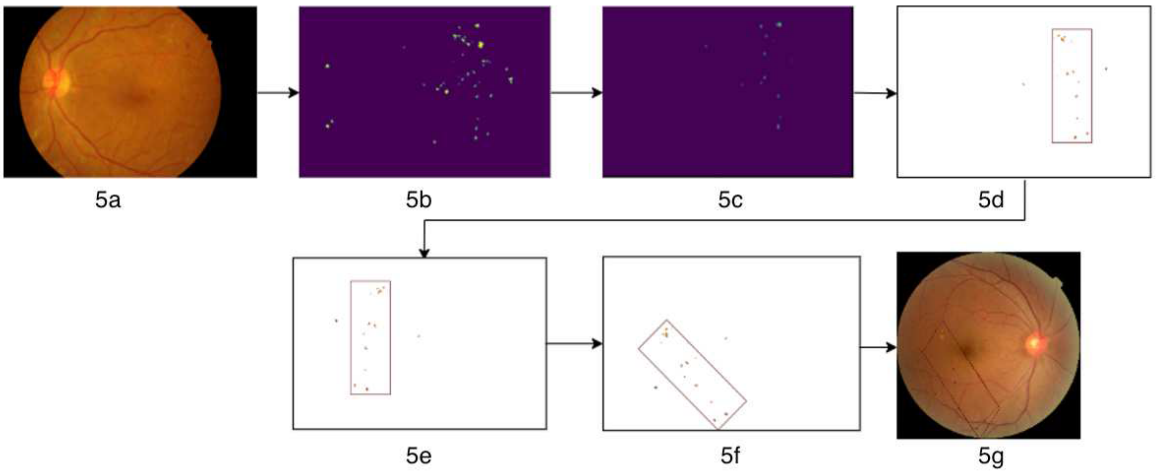

Figure 6.5: A detailed step-by-step illustration of the proposed algorithm.

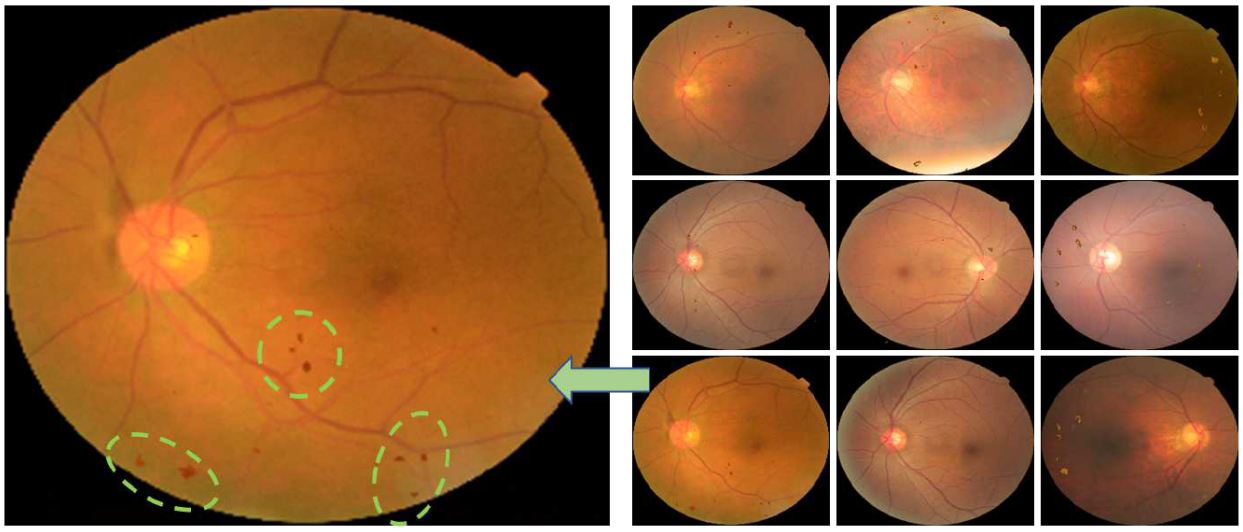

Figure 6.6: Synthetic data after projecting lesions onto healthy images (right side); magnified view of one sample case with some lesions annotated (left side).

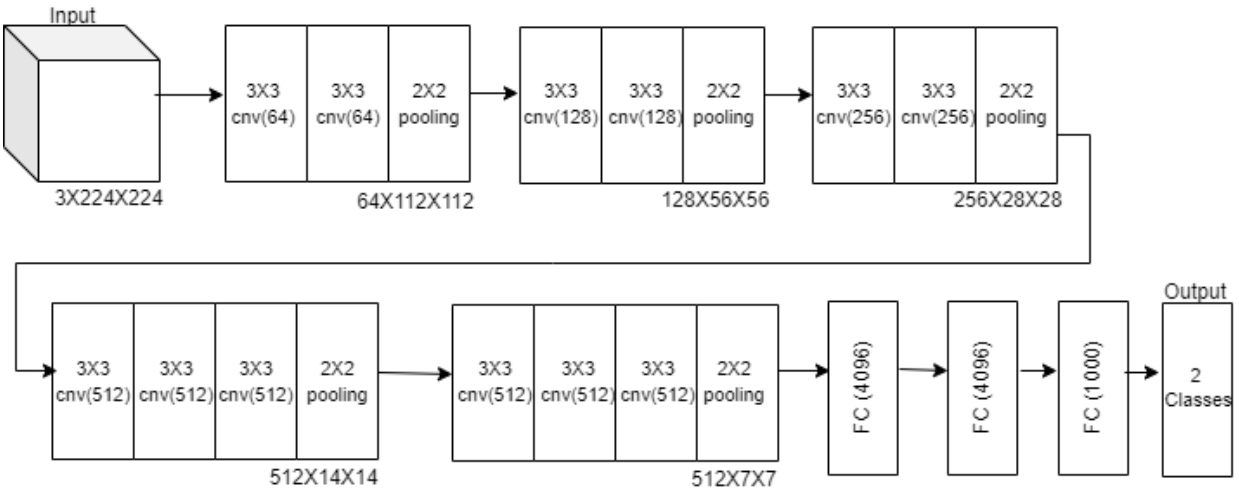

Figure 6.7: Illustration of the modified VGG-16 network.

many DCNN architrectures have been developed and used as transfer learning models in medical imaging informatics, the most popular DCNN architrectures, VGG16, ResNet-50, and Inception-v3 architectures are used.

Two VGG16-network-based deep transfer learning 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. Figure 6.7 shows the modified VGG-16 architecture. Specifically, 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. Since the VGG16 architecture is pre-trained on the ImageNet database and accepts input images with a size of 225 $\times$ 225 pixels, all retinal fundus images were resized to be 225 $\times$ 225 pixels. Next, two ResNet-50 network-based deep learning models and two Inception-v3 network-based deep learning models were trained using original and synthetic data, respectively. Figure 6.8 and Figure 6.9 show the deep learning architecture of the modified ResNet-50 network and the modified Inception-v3 network, respectively.

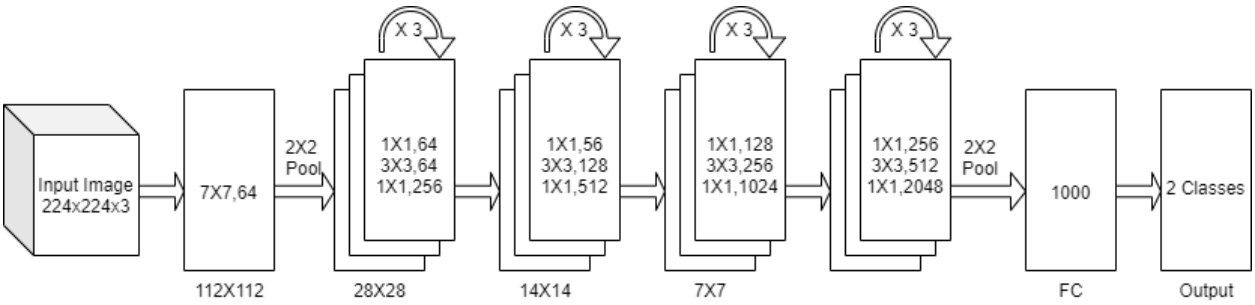

Figure 6.8: Illustration of the modified ResNet-50 network.

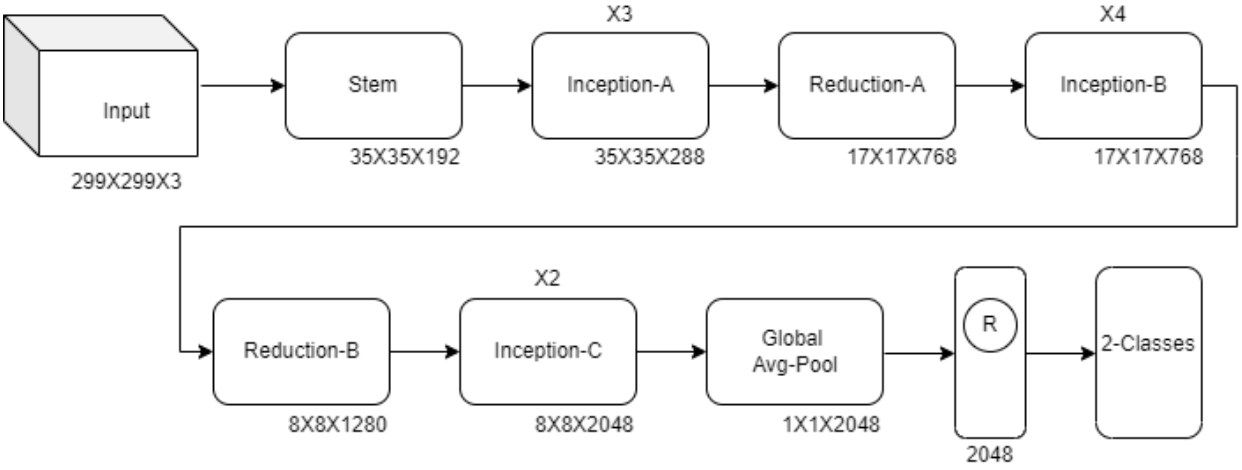

Figure 6.9: Illustration of the modified Inception-v3 network

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 and changed to one output neuron with sigmoid activation that achieves the goal of classifying two categories of retinal fundus images in this study.

After fine-tuning and optimizing each DCNN model, each model was applied to the same independent test dataset of 27 images in the IDRiD dataset. 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 with the higher probability score. From the test results, true positive (TP), false negative (FN), true negative (TN), and false positive (FP) values to generate a confusion matrix are generated. The model is evaluated for classification performance from the confusion matrix by computing four commonly used evaluation indices: accuracy, precision, recall, and F1 score, using Equations (Refer to appendix 9).

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

6.4 Results

Figures 6.10 and 6.11 shows six diagrams that plot curves of disease classification accuracy over 25 training epochs of the three deep transfer learning models (VGG16, ResNet-50, and Inception-v3) trained using the original IDRiD images and the synthetic images, respectively.

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 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. The model has 25 training epochs for all six models.

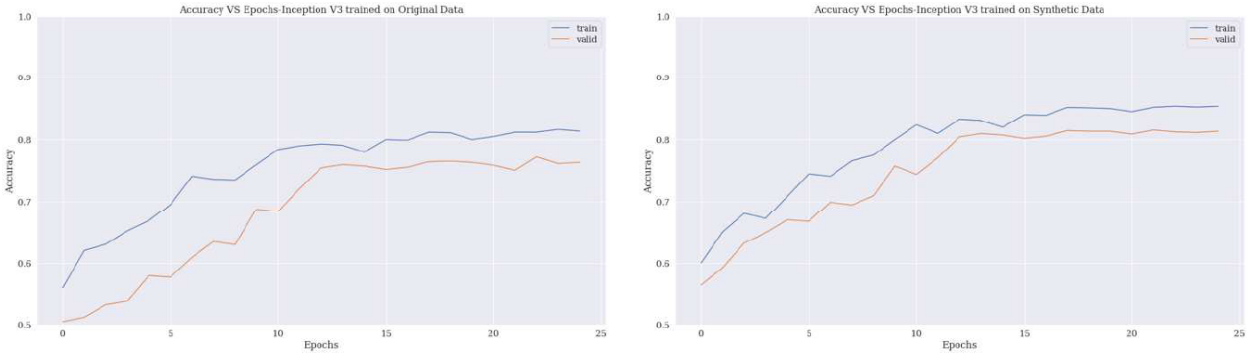

Figure 6.10: Training vs validation accuracy curves of Inception-v3 model using the original IDRiD images (left) and the synthetic images (right)

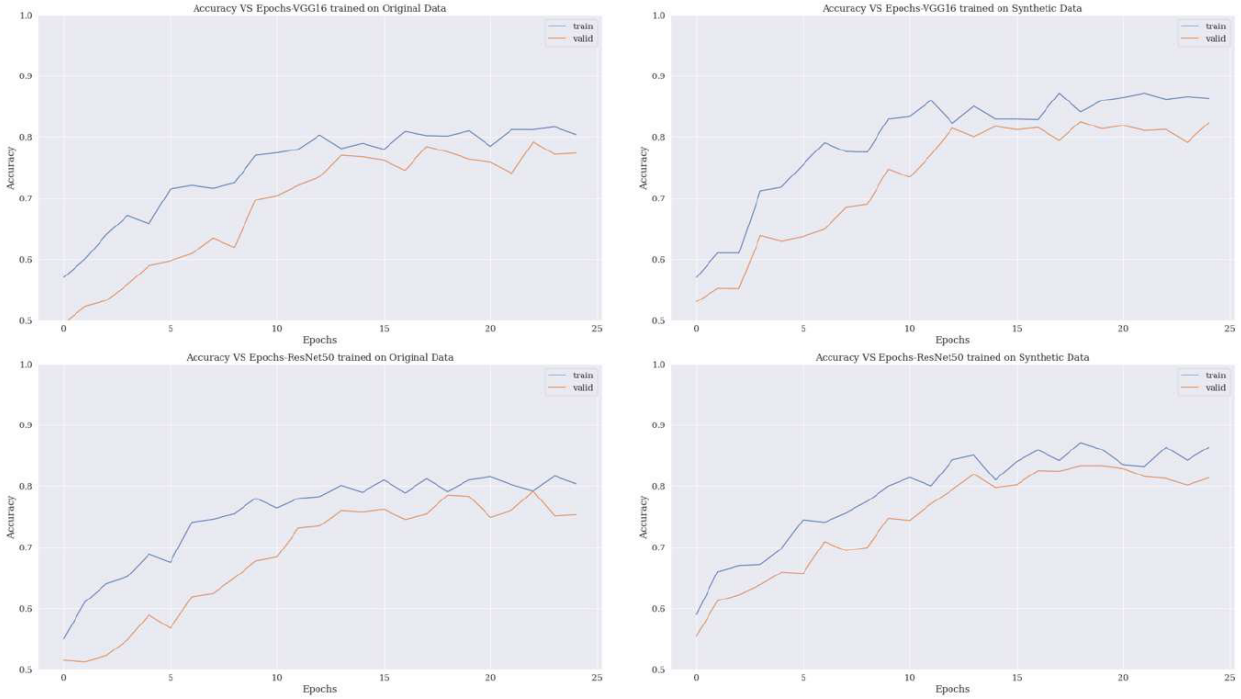

Figure 6.11: Training and validation accuracy curves of VGG16(top) and ResNet-50(bottom) model using the original IDRiD images (left) and the synthetic images (right)

The testing dataset includes 14 images in Category 1 which includes Hemorrhages, Hard Exudates, and Soft Exudates, and 13 images in Category 2 that includes Hemorrhages and Hard Exudates. Figure 6.12 and 6.13 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 6.2 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.

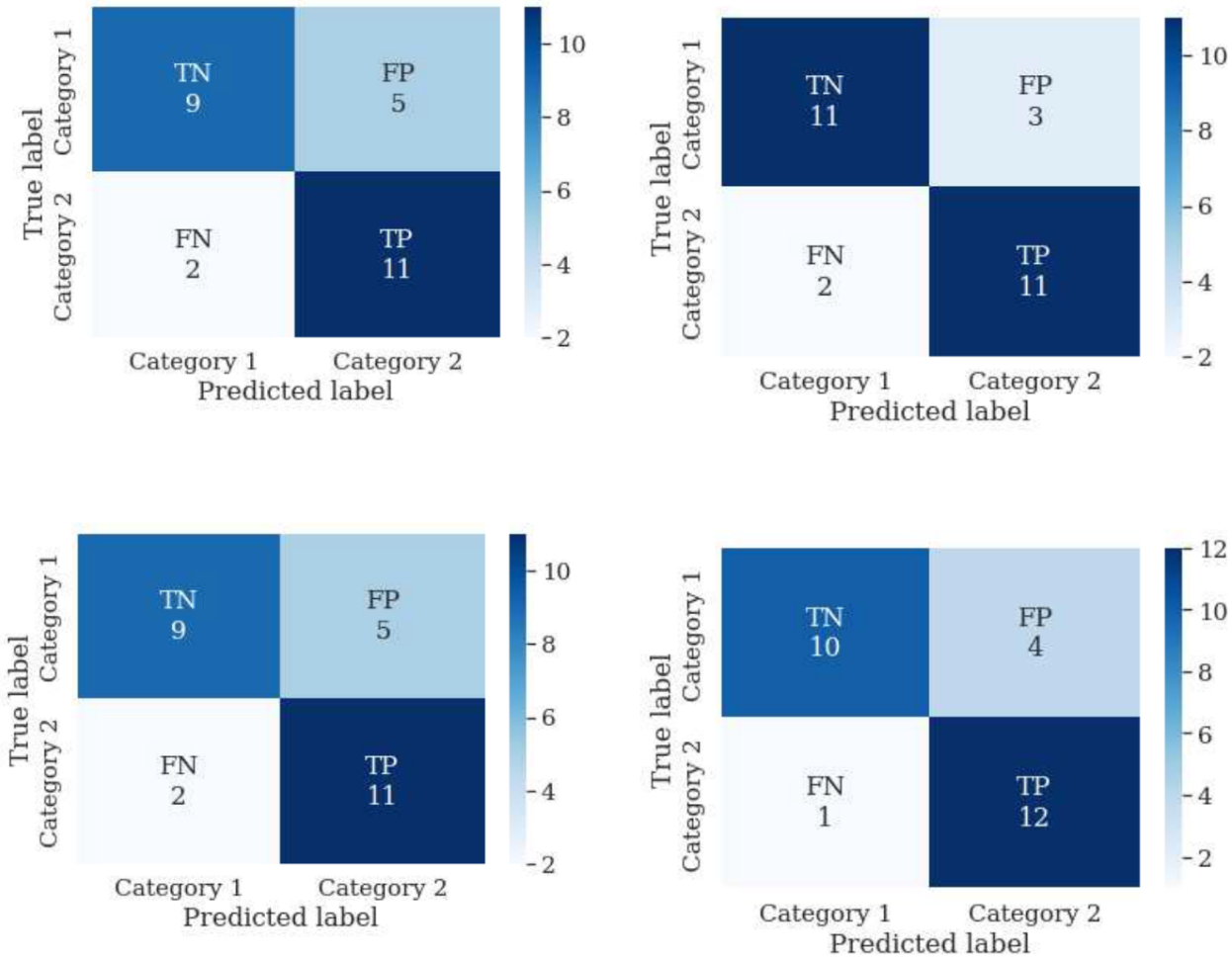

Figure 6.12: Confusion matrices generated by VGG16 (top), ResNet-50 (bottom) trained using original IDRiD images (left) and synthetic images (right).

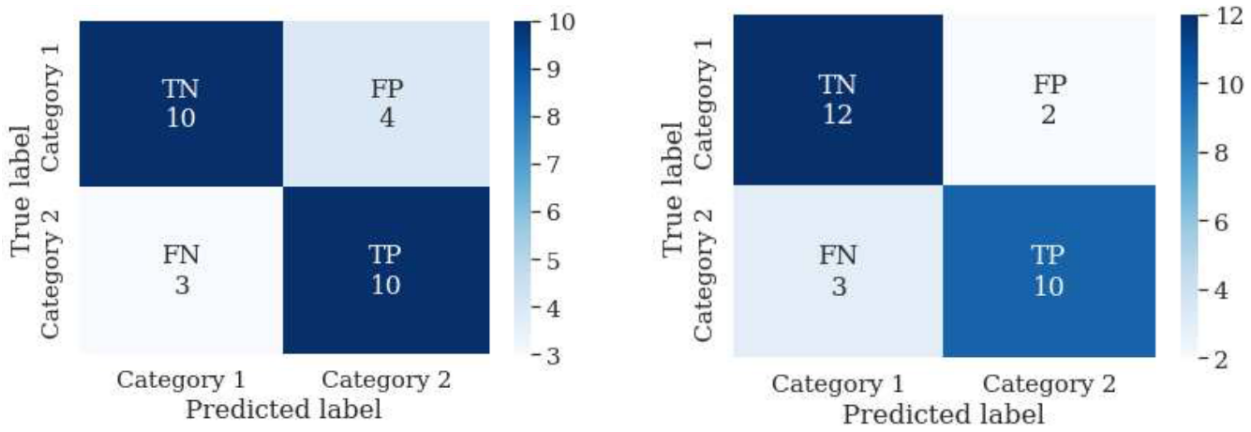

Figure 6.13: Confusion matrices generated by Inception-v3 trained using original IDRiD images (left) and synthetic images (right).

Model	Training Images	Accuracy	Precision	Recall	F1 Score
VGG-16	Original	0.741	0.714	0.769	0.741
VGG-16	Synthetic	0.815	0.833	0.769	0.800
ResNet-50	Original	0.741	0.688	0.846	0.759
ResNet-50	Synthetic	0.815	0.786	0.849	0.815
Inception-v3	Original	0.741	0.698	0.847	0.759
Inception-v3	Synthetic	0.815	0.750	0.923	0.828

Table 6.2: Comparison of various performance metrics while using original and synthetic data

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) differ. The transfer learning models fine-tuned using the original IDRiD images yield substantially higher precision as the models fine-tuned using the synthetic images. The results indicate that fine-tuned synthetic image models 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).

6.5 Discussion

Due to the difficulty of acquiring large datasets of well-annotated medical images for medical imaging research, generating reliable and diverse synthetic images plays an important role in improving the efficacy of either building new deep learning models or fine-tuning the existing deep-learning models for medical imaging applications. Thus, there is broad interest in medical imaging research to develop more efficient and robust algorithms to generate synthetic image data with high clinical relevance. This study proposes and demonstrates a new algorithm to generate synthetic retinal fundus images embedded with different types of diseased lesions. The new synthetic image data generation method has several unique characteristics, and the study results also generate several interesting observations.

First, current existing synthetic image data generation methods or algorithms include a Monte Carlo method based on repeated random sampling and statistical analysis of the results, an unsupervised Variational Auto-Encoder (VAE)(23) algorithm that learns the distribution of the original image dataset and generate synthetic image data via double transformation using an encoded-decoded architecture, and a Generative Adversarial Network (GAN) (15) that uses two neural networks working together to generate fake yet realistic data points. These methods are quite complicated to design, train, and implement, and are computationally expensive because large numbers of algorithm or network parameters need to be chosen and optimized. However, our new method is much simpler and computationally efficient. It uses a multi-stage approach to extract positive lesions directly, randomize distribution of lesion blobs, and then inserts or projects the 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 manually annotated diseased regions or lesions. This study demonstrates that by using this algorithm, we can

significantly expand our dataset size from 54 original positive images and 60 negative or normal images to 14,994 synthetic images (for Categories 1 and 2 diseases as shown in Table 6.1).

Second, although the original 54 images are divided into only two categories, using this new algorithm, we can generate all varieties of patient data (Categories 1-7) as shown in Table 6.1 In this study, the algorithm generates 49,920 synthetic images in 7 categories. Please note that despite the small image base of 54 positive images and 60 negative images, and large number of synthetic images, all algorithm-generated synthetic images have different positive lesion blob combinations. The lesion blobs randomly distribute in different locations with different orientations, which increases diversity of training samples by avoiding or minimizing redundancy of the generated synthetic image data.

Third, three sets of DCNN models based VGG16, ResNet 50 and Inception V3 model are used. Each DCNN set includes two models. Model 1 is trained using original IDRiD images and Model 2 is trained using the algorithmgenerated synthetic image data. Two models are then applied to the same testing images. The experimental results show that Model 2 yields higher classification accuracy than Model 1. Specifically, among 27 testing cases, Model 1 correctly classifies 20 cases, while using Model 2 correctly classifies 22 cases with a classification accuracy increase by 7.4%. This accuracy improvement is observed on a quite small testing image dataset of 27 images. The two models should be further tested using much larger image datasets. However, the results still demonstrate the feasibility and advantages of using new simple algorithm to generate synthetic images to substantially increase size of training dataset and potentially increase model classification performance on the new independent testing images.

Fourth, It is observed that classification performance as measured by 4 evaluation indices (as shown in Table 6.2) is independent from three deep learning DCNN models (VGG-16,

ResNet 50 and Inception v3). It is promising that all three types of different 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) as shown in Table 6.2. This observation indicates the robustness or scientific rigor of applying this new simple multi-step algorithm to generate synthetic images that can effectively help train or fine-tune different DCNN models.

Finally, although study results are promising, we also recognize that seamless insertion of lesions or other abnormality regions onto negative images using this simple algorithm has restrictions or limitations, which include that (1) the lesions must have clear boundary contours so that the lesions can be easily segmented or 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 or irregular boundary embedded under heterogenous tissue background (i.e., breast tumors depicting on mammograms), this algorithm will not work âas isâ and modifications will be needed. Despite such limitations, developing this new simple algorithm to generate synthetic images still has its higher clinical application potential or impact at least for retinal fundus images that are acquired using a low-cost image examination method and thus widely used in clinics to screen, detect, and diagnose many common human diseases including a variety of eye diseases and diabetes. We will further test and validate this new algorithm and apply it to develop more accurate and robust deep learning models for different medical applications in the future.

Chapter 7

Conclusions

In conclusion, the use of hybrid computer vision algorithms has proven to be a promising approach to improving various tasks in medical image analysis, such as image segmentation, classification, detection, and synthetic data generation.

Hybrid algorithms combine the advantages of multiple techniques to address the limitations of individual algorithms. For instance, a hybrid algorithm may combine the strengths of both deep learning and traditional image processing methods to achieve better segmentation accuracy. Multi-stage algorithms, on the other hand, break down complex tasks into smaller, more manageable sub-tasks, resulting in more efficient and accurate results.

Combining multiple algorithms can also improve image classification and detection tasks, as it can incorporate various features and contexts into the analysis. For instance, a multi-stage algorithm may extract features and then classify them to detect abnormalities in medical images.

Moreover, hybrid or multi-stage algorithms can facilitate the generation of synthetic medical data, which can be used to address the scarcity of labeled data in medical image analysis. Artificial data generation algorithms, such as GANs, can generate realistic medical images

with variations in anatomical structures and image appearances, improving the performance of deep learning models.

In summary, using hybrid or multi-stage computer vision algorithms has demonstrated its potential to improve the accuracy, efficiency, and efficacy of medical image analysis tasks, making it a promising area for future research and development.

Chapter 8

Future Work

This chapter discusses challenges and possible future advancements to the proposed models to address these issues.

8.1 Problem: Interpretability of deep learning features

Interpretability of deep learning model features is crucial in medical image analysis. While deep learning models, such as convolutional neural networks (CNNs), can achieve high accuracy in image classification and segmentation tasks, the complex internal workings of these models can make it difficult to interpret the features they learn. Here are some of the main issues related to the interpretability of deep learning model features in medical image analysis:

Black Box Nature: Deep learning models are often described as "black box" models because the internal workings of the model are not easily interpretable by humans. The complex interplay between the layers of a CNN can make it difficult to understand how

the model is making its predictions. This can be problematic in medical image analysis, where it is important to understand the underlying features that contribute to a diagnosis or segmentation.

Lack of Transparency: In addition to the black box nature of deep learning models, there is often a lack of transparency in how the model was trained and how it arrived at a particular prediction. It can be difficult to determine which features the model is using to make its predictions, or whether those features are clinically relevant.

Data-Driven Features: Deep learning models learn features from the data itself, rather than relying on pre-defined features as in traditional machine learning. While this can lead to improved accuracy, it can also make it difficult to understand the clinical relevance of the features the model has learned.

Non-intuitive Features: The features learned by deep learning models may not be immediately interpretable by humans. For example, a CNN may learn a particular texture pattern in an image that is difficult for humans to identify or understand. This can make it challenging to understand how the model is making its predictions and to validate the clinical relevance of the features.

Domain-Specific Knowledge: Medical image analysis requires a deep understanding of the underlying anatomy and physiology of the human body. Deep learning models may not capture this domain-specific knowledge, making it difficult to interpret the features the model has learned in the context of clinical practice.

Limited Validation: While deep learning models may achieve high accuracy on validation datasets, it can be difficult to validate the clinical relevance of the features they have learned. This is particularly true for rare or complex medical conditions, where there may not be enough data to validate the features learned by the model.

Addressing these challenges is crucial for improving deep learning models' clinical rele-

vance and applicability in medical image analysis.

8.2 Possible Solution: Combination of traditional and deep learning features for better interpretability

One possible future direction for research using hybrid computer vision models to address the interpretability of deep learning features compared with traditional models could involve integrating radiomics feature extraction into the model architecture. *Radiomics* is an approach that involves extracting quantitative features from medical images to aid in diagnosis, prognosis, and treatment planning.

The integration of radiomics feature extraction into hybrid models could help to improve the interpretability of deep learning features by providing additional context and clinical relevance to the analysis. For instance, the hybrid model could first extract radiomic features from the medical images, which could then be used to guide the deep learning model's feature selection and classification process. This could help identify the most relevant and clinically significant features that contribute to the model's decision-making process, making it more interpretable.

Moreover, the combination of radiomics and deep learning models could also help to address the limitations of individual approaches. For instance, radiomics-based models may be limited by discriminative features, while deep learning models may be limited by the lack of interpretability. By combining these approaches, the hybrid model could leverage the strengths of both approaches to improve the accuracy and interpretability of medical image analysis.

Furthermore, integrating radiomics feature extraction into hybrid models could help to facilitate the development of personalized medicine approaches in medical image analysis.

The hybrid model could extract patient-specific radiomics features, which could be used to tailor the analysis to the patient's specific clinical characteristics, such as age, gender, and disease stage.

Overall, integrating radiomics feature extraction into hybrid computer vision models could represent a promising approach to address the interpretability of deep learning features compared with traditional models, leading to more accurate and clinically relevant medical image analysis.

Chapter 9

Appendix

Precision: Precision is a measure of the fraction of true positive pixels (correctly classified as positive) out of all pixels predicted as positive by the model. Precision can be calculated as:

$$Precision = TP/(TP + FP) \tag{9.1}$$

where TP is the number of true positive pixels and FP is the number of false positive pixels.

Accuracy: Accuracy is a measure of the fraction of correctly classified pixels out of all pixels in the image. Accuracy can be calculated as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{9.2}$$

where TN is the number of true negative pixels and FN is the number of false negative pixels.

Dice Coefficient: Dice coefficient is a measure of the overlap between the predicted segmentation mask and the ground truth segmentation mask. The Dice coefficient ranges

from 0 (no overlap) to 1 (perfect overlap). The Dice coefficient can be calculated as:

$$Dice = 2 * TP/(2 * TP + FP + FN) \tag{9.3}$$

Intersection over Union (IoU) score: IoU is another measure of the overlap between the predicted segmentation mask and the ground truth segmentation mask. The IoU score ranges from 0 (no overlap) to 1 (perfect overlap). IoU can be calculated as:

$$IoU = TP/(TP + FP + FN) \tag{9.4}$$

where TP is the number of true positive pixels, FP is the number of false positive pixels, and FN is the number of false negative pixels.

Recall: Recall measures the ability of a model to identify all relevant instances in the dataset. It is the ratio of the true positive (TP) predictions to the sum of TP and false negative (FN) predictions. In image classification, a high recall score indicates that the model is able to correctly identify most of the relevant objects or classes present in the images.

$$Recall = TP/(TP + FN) \tag{9.5}$$

F1 Score: F1 score is a weighted harmonic mean of precision and recall. It is a balance between precision and recall, and considers both false positives (FP) and false negatives (FN) in the evaluation of the model's performance. In image classification, a high F1 score indicates that the model is able to correctly identify both relevant and non-relevant objects or classes present in the images.

$$F1Score = 2 * (precision * recall)/(precision + recall) \tag{9.6}$$

Bibliography

[1] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[2] Hidetaka Arimura, Shigehiko Katsuragawa, Kenji Suzuki, Feng Li, Junji Shiraishi, Shusuke Sone, and Kunio Doi. Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening1. *Academic Radiology*, 11(6):617–629, 2004.

[3] Y Bar, I Diamant, L Wolf, and et al. Chest pathology detection using deep learning with non-medical training. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 294–297, 2015.

[4] Chi-Ming Chen, Yi-Hong Chou, Noriko Tagawa, and Yi Do. Computer-aided detection and diagnosis in medical imaging. *Computational and Mathematical Methods in Medicine*, 2013:790608, 2013.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Yibin Chen, Mingyang Wu, Ruiping Tang, Shuai Chen, and Senbo Chen. A hybrid deep learning model based on lstm for long-term pm2. 5 prediction. In *2021 the 3rd International Conference On Intelligent Science And Technology (ICIST)*, pages 55–60, 2021.

[8] Gopi Danala, Sai Kiran Maryada, Morteza Heidari, and et al. A new interactive visual-aided decision-making supporting tool to predict severity of acute ischemic stroke. In *Proceedings of SPIE*, volume 11317, page 113171V, 2020.

[9] Gopi Danala, Sai Kiran Maryada, Wakil Islam, Morteza Heidari, and et al. A comparison of computer-aided diagnosis schemes optimized using radiomics and deep transfer learning methods. *Bioengineering*, 9(6):256, 2022.

[10] Gopi Danala, Seyedehnazanin Mirniaharikandehei, Mark Jones, and et al. Developing interactive computer-aided detection tools to support translational clinical research. In *Proceedings of SPIE*, volume 12035, page 1203503, 2022.

[11] Cach N Dang, María N Moreno-García, and Fernando De la Prieta. Hybrid deep learning models for sentiment analysis. *Complexity*, 2021:1–16, 2021.

[12] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[13] M Frid-Adar, I Diamant, E Klang, and et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[16] Kaiming He, Georgia Gkioxari, Piotr DollÃ¡r, and Ross Girshick. Mask r-cnn, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Mohammad Asiful Hossain, Rezaul Karim, Ruppa Thulasiram, Neil DB Bruce, and Yang Wang. Hybrid deep learning model for stock price prediction. In *2018 ieee symposium series on computational intelligence (ssci)*, pages 1837–1844. IEEE, 2018.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] Muhammad Hussain, Jordan J Bird, and Diego R Faria. A study on cnn transfer learning for image classification. In *Advances in Intelligent Systems and Computing*, volume 840, pages 191–202. Springer, 2019.

[21] Matthew A Jones, Wasiq Islam, Rashid Faiz, and et al. Applying artificial intelligence technology to assist with breast cancer diagnosis and prognosis prediction. *Frontiers in Oncology*, 12:980793, 2022.

[22] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[23] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[24] S Kiran Maryada et al. An efficient synthetic data generation algorithm to improve efficacy of deep learning models of medical images. *Researchsquare.com*, 2022.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, page 3065386, 2012.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[29] Momina Masood, Tahira Nazir, Marriam Nawaz, Awais Mehmood, Junaid Rashid, Hyuk-Yoon Kwon, Toqeer Mahmood, and Amir Hussain. A novel deep learning method for recognition and classification of brain tumors from mri images. *Diagnostics*, 11(5):744, 2021.

[30] Timothy McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2):91–108, 1996.

[31] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, and et al. Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. In *Medical Image Computing and Computer-Assisted Intervention â MICCAI*, pages 893–901, 2018.

[32] Louis A Meinel, Alan H Stolpen, Kevin S Berbaum, and et al. Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system. *Journal of Magnetic Resonance Imaging*, 25(1):89–95, 2007.

[33] Sanjana Mudduluru, Sai Kiran Reddy Maryada, William Lee Booker, Dean F. Hougen, and Bin Zheng. Improving medical image segmentation and classification using a novel joint deep learning model. In Khan M. Iftekharuddin and Weijie Chen, editors, *Medical Imaging 2023: Computer-Aided Diagnosis*, volume 12465, page 124652H. International Society for Optics and Photonics, SPIE, 2023.

[34] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[35] Dzung L Pham, Chunming Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2:315–337, 2000.

[36] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[38] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, mar 2020.

[39] HC Shin, K Roberts, L Lu, and et al. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Narotam Singh, Neha Soni, Amita Kapoor, et al. Automated detection of alzheimer disease using mri images and deep neural networks-a review. *arXiv preprint arXiv:2209.11282*, 2022.

[42] William Song and Jim Cai. End-to-end deep neural network for automatic speech recognition. *Standford CS224D Reports*, pages 1–8, 2015.

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[45] Qi Tang, Guoen Xia, Xianquan Zhang, and Feng Long. A customer churn prediction model based on xgboost and mlp. In *2020 International Conference on Computer Engineering and Application (ICCEA)*, pages 608–612. IEEE, 2020.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[47] Lulu Wang. Deep learning techniques to diagnose lung cancer. *Cancers*, 14(22):5569, 2022.

[48] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[49] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[50] Yiming Zhang and Chong Wang. Siim-isic melanoma classification with densenet. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 14–17. IEEE, 2021.

[51] Yefeng Zheng, Kelvin KL Wong, Xuejun Gu, Qinghua Zhang, and Stephen T Wong. Automatic aorta segmentation and valve landmark detection in c-arm ct for transcatheter aortic valve implantation. *IEEE Transactions on Medical Imaging*, 31(12):2307–2321, 2012.