Master Thesis: An Ensemble Deep Transfer Learning Model to Enhance Diabetic Retinopathy Diagnosis in Retina Images

Chapter 1: Introduction

Context and Motivation
Diabetic retinopathy (DR) is one of the leading causes of blindness among the working-age population worldwide. As a complication of diabetes, it progressively damages the retina, often without noticeable symptoms until advanced stages. Early detection is critical, as timely intervention can prevent up to 95% of vision loss cases. However, the manual grading of retinal images by ophthalmologists is time-consuming, expensive, and prone to inter-observer variability — especially in regions with limited access to specialists.

With the rapid development of artificial intelligence and deep learning, computer-aided diagnosis systems have become increasingly viable and accurate. Deep convolutional neural networks (CNNs), transformers, and advanced ensemble methods now offer state-of-the-art performance in medical imaging tasks, including diabetic retinopathy detection.

Problem Statement
Despite recent progress, the accurate and reliable classification of diabetic retinopathy into its five severity stages remains a challenge due to:
- The subtle visual differences between adjacent stages
- Limited availability of labeled medical data
- The presence of image quality issues (e.g., blur, brightness)

The goal of this thesis is to design an ensemble deep transfer learning model capable of enhancing diabetic retinopathy diagnosis accuracy and robustness using retina images from the APTOS 2019 dataset.

Objectives

The main objectives of this research are:

1. To apply and fine-tune multiple state-of-the-art deep learning models (ResNet50, EfficientNetB0, VGG16, Vision Transformer, and CapsNet) using transfer learning.
2. To evaluate and compare their performance on the task of multi-class diabetic retinopathy classification.
3. To design an ensemble strategy that combines these models to improve overall performance and reduce classification errors.
4. To analyze the effectiveness of this approach and identify key factors influencing model performance.

Structure of the Thesis

This thesis is organized into six chapters:

- Chapter 1 presents the general introduction, motivation, and research objectives.
- Chapter 2 reviews the scientific background and related work in diabetic retinopathy diagnosis and deep learning.
- Chapter 3 describes the methodology, including data preprocessing, model architectures, and training strategies.
- Chapter 4 details the implementation setup and experiments conducted.
- Chapter 5 presents the results, evaluation metrics, and discussion.
- Chapter 6 concludes the thesis and outlines possible directions for future work.

Chapter 2: Literature Review

2.1 Diabetic Retinopathy: Overview

Diabetic retinopathy (DR) is a diabetes-induced complication that affects the small blood vessels of the retina. It is characterized by microaneurysms, hemorrhages, exudates, and, in advanced stages, neovascularization. DR progresses through five stages:

1. No DR
2. Mild non-proliferative DR
3. Moderate non-proliferative DR
4. Severe non-proliferative DR
5. Proliferative DR

Traditional diagnosis is performed manually by ophthalmologists through fundus

image examination. However, this process is resource-intensive and often inconsistent across practitioners.

2.2 Deep Learning in Medical Imaging

Deep learning has revolutionized medical image analysis, particularly with the advent of Convolutional Neural Networks (CNNs). CNNs can automatically learn spatial hierarchies of features directly from image data. Notable applications include:

- Skin lesion classification
- Pneumonia detection from chest X-rays
- Brain tumor segmentation in MRIs
- Diabetic retinopathy classification from fundus images

These models significantly outperform traditional machine learning algorithms that rely on hand-crafted features.

2.3 Transfer Learning

Transfer learning involves reusing pre-trained models — typically trained on large datasets like ImageNet — and fine-tuning them for a specific task. This approach is especially effective in medical domains, where annotated data is often scarce. Popular transfer learning models include:

- ResNet50: Deep residual learning with skip connections to mitigate vanishing gradients.
- VGG16: Known for its simplicity and uniform architecture, although computationally heavier.
- EfficientNetB0: Combines network depth, width, and resolution scaling for high efficiency and accuracy.
- Vision Transformer (ViT): Applies self-attention mechanisms instead of convolution, enabling global context modeling.
- CapsNet: Captures spatial relationships through capsule units, offering better viewpoint invariance.

2.4 Ensemble Learning in Deep Learning

Ensemble learning aggregates predictions from multiple models to produce more robust and accurate outcomes. Techniques include:

- Hard Voting: Majority class decision from multiple classifiers.
- Soft Voting: Average of predicted class probabilities.
- Stacking: Meta-models are trained on outputs of base models.

In medical imaging, ensembles have shown better generalization and reduced overfitting, especially when models have complementary strengths.

2.5 Related Work

Several studies have addressed diabetic retinopathy classification:
- Gulshan et al. (2016) used a deep CNN to classify DR with sensitivity >90% on EyePACS.
- Pratt et al. (2016) employed a modified CNN achieving 75% accuracy on Kaggle data.
- A recent ensemble approach by Porwal et al. (2020) combined VGG, ResNet, and DenseNet to improve diagnosis reliability.
- Vision transformers have also shown promise in DR tasks by modeling long-range dependencies.

However, many studies are limited to binary classification (DR vs. No DR) or use small datasets. Our work extends this by applying and combining multiple deep models for multi-class classification, with a focus on ensemble-based enhancement.

2.6 Summary

This chapter has reviewed the progression of diabetic retinopathy, the rise of deep learning in medical imaging, and key methods like transfer and ensemble learning. Despite the success of individual models, their combination in an ensemble has been underexplored for fine-grained DR classification. Our work builds on this gap to propose a powerful ensemble of deep transfer models for improved diagnosis.

Chapter 3: Methodology

3.1 Dataset Description

The dataset used for this study is the APTOS 2019 Blindness Detection dataset, available on Kaggle. It consists of high-resolution retina images labeled into five categories representing the severity of diabetic retinopathy:

- 0: No DR
- 1: Mild
- 2: Moderate
- 3: Severe
- 4: Proliferative DR

The dataset includes 3,662 training images with corresponding labels, and an unlabelled test set used for Kaggle evaluation.

3.2 Data Preprocessing

Before feeding the images into deep learning models, we applied the following preprocessing steps:

- Image resizing to 224×224 (or 299×299 for certain models like Inception)
- Normalization of pixel values to the [0, 1] range
- Data augmentation to reduce overfitting:
  - Horizontal/vertical flipping
  - Rotation (±15°)
  - Brightness and contrast adjustment
- Shuffling and stratified splitting into training and validation sets (e.g. 80/20 split)

3.3 Model Architectures

We implemented and fine-tuned the following pre-trained models:

- ResNet50: A residual network using skip connections to improve gradient flow in deep layers.
- EfficientNetB0: A lightweight model balancing depth, width, and resolution, known for high efficiency.
- VGG16: A deep convolutional network with a simple and uniform architecture.
- Vision Transformer (ViT): A transformer-based model that uses attention mechanisms to model image patches.
- CapsNet: Capsule networks that preserve spatial hierarchies between image features.

Each model was initialized with ImageNet weights and fine-tuned on the APTOS dataset.

3.4 Ensemble Strategy

To enhance the prediction performance, we built an ensemble model combining the outputs of the five base models. We experimented with:

- Soft voting: Averaging the predicted probabilities from all models.
- Hard voting: Taking the majority class from the individual model predictions.
- In some configurations, weighted averaging was applied, assigning more importance to the best-performing models.

3.5 Training Configuration

- Framework: TensorFlow + Keras (Google Colab environment)
- Loss function: Categorical Crossentropy
- Optimizer: Adam (initial LR = 0.0001)
- Batch size: 32
- Epochs: 20–50, with EarlyStopping and ModelCheckpoint
- Evaluation metrics: Accuracy, F1-score, Confusion Matrix, AUC