A Resource-Efficient Deep Learning Framework for Gastrointestinal Image Classification

BY

Sowmik Hasan Niloy
ID: 242-25-009

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. Arif Mahmud
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor & Director, M.Sc in CSE
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

APPROVAL

This Project/Thesis titled **"A Resource-Efficient Deep Learning Framework for Gastrointestinal Image Classification"**, submitted by **Sowmik Hasan Niloy**, ID No: **242-25-009** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Nazibur Rahman
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. Arif Mahmud, Associate Professor and Associate Head,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Arif Mahmud
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor & Director, M.Sc in CSE
Department of CSE
Daffodil International University

Submitted by:

Sowmik Hasan Niloy
ID: 242-25-009
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to Almighty Allah for His divine blessing, which makes it possible to complete the final year project/internship successfully.

I am grateful and wish to express my profound indebtedness to **Dr. Arif Mahmud, Associate Professor and Associate Head,** Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the** Department of CSE, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

In this paper, we propose a cost-effective deep learning solution for diagnosing GI diseases, focusing on GERD and colorectal polyps, through endoscopic images. We explore the problems of executing high-performance AI models in resource-restricted clinical settings through incorporating transfer learning, ensemble learning, and a new lightweight architecture named MiniMedNet (~32.5k parameters). The dataset is a global dataset and includes four classes (GERD, GERD Normal, Polyp, Polyp Normal) which are collected from one publicly available Mendeley repository and heavily pre-processed (augmentation, resizing, normalization). Several off-the shelf models (EfficientNetB3, ResNet50, DenseNet121, and MobileNetV2) were tested individually and in ensemble for the purpose of setting the baselines. We show that MiniMedNet, a network model designed from the ground up to use the fewest number of parameters possible, attains a test accuracy of 78\% on average, which is comparable to the performance of other more compute-laden models. Exhaustive analysis with metrics accuracy, precision, recall, F1-score and Grad-CAM visualizations reiterate the ability of the model to retain interpretatibility and reliability. Our approach presents a potential route for deploying AI-supported diagnosis in resource-limited healthcare, by trading-off between diagnostic performance and computational cost, and by enabling scalable, clinically practical AI deployment.

Keywords: Gastroesophageal Reflux Disease, Colorectal Polyps, Deep Learning, Transfer Learning, Ensemble Learning, Lightweight CNN, MiniMedNet, Medical Imaging AI

.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-3**
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Research Objectives	2
1.4 Research Questions	2
1.5 Expected Output	2-3
1.6 Project Management and Finance	3
1.7 Report Layout	3
CHAPTER 2: BACKGROUND	**4-9**
2.1 Preliminaries	4
2.2 Related Works	4-8
2.3 Research Gap	8-9
2.4 Challenges	9
CHAPTER 3: RESEARCH METHODOLOGY	**10-22**
3.1 Proposed Methodology	10-13
3.2 Data Collection Procedure	14-15
3.3 Image pre-processing	16
3.3.1 Resizing Images	16-117
3.3.2 Normalization to[0,1] Range	17
3.3.3 DataAugmentation	17
3.3.4 Class-Balanced Stratified Splitting	18
3.3.5 Batch Preparation and Shuffling	18
3.3.6 Preprocessing Pipeline Compatibility	18-19
3.4 Deep Learning Models	19

3.4.1 MiniMedNet	19
3.4.2 EfficientNetB3	20
3.4.3 ResNet50	21
3.4.4 DenseNet121	22
3.4.5 MobileNetV2	22
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	24-41
4.1 Results of Grad-CAM Visualization	24-28
4.2 Evolution Methods	28-29
4.3 Experimental Results & Analysis	29-39
4.4 Discussion	39-41
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	42-44
5.1 Impact on Society	42
5.2 Impact on Environment	43
5.3 Ethical Aspects	43-44
5.4 Sustainability Plan	44
CHAPTER 6: CONCLUSION AND FUTURE WORK	45-46
6.1 Summary of the Study	45
6.2 Conclusions	45-46
6.3 Implication for Further Study	46
REFERENCES	**47-50**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Proposed methodology for gastrointestinal disease classification.	11
Figure 3.2: Random samples per class from the gastrointestinal endoscopy dataset used in this study.	15
Figure 3.3 Architecture of MiniMedNet	20
Figure 3.4 Architecture of EfficientNetB3	21
Figure 3.5 Architecture of ResNet50	21
Figure 3.6 Architecture of DenseNet121	22
Figure 3.7 Architecture of MobileNetV2	23
Figure 4.1 Grad-CAM visualizations of EfficientNetB3	24
Figure 4.2: Grad- CAM visualizations of DenseNet121	25
Figure 4.3: Grad- CAM visualizations of ResNet50	25
Figure 4.4: Grad- CAM visualizations of MobileNetV2	26
Figure 4.5: Grad- CAM visualizations of MiniMedNet	26
Figure 4.6: Accuracy and Loss Curves of EfficientNetB3	31
Figure 4.7: Confusion Matrix of EfficientNetB3	32
Figure 4.8: Accuracy and Loss Curves of ResNet50	32
Figure 4.9: Confusion Matrix of EfficientNetB3	33
Figure 4.10: Accuracy and Loss Curves of DenseNet121	33
Figure 4.11: Confusion Matrix of DenseNet121	34
Figure 4.12: Accuracy and Loss Curves of MobileNetV2	34
Figure 4.13: Confusion Matrix of MobileNetV2	35
Figure 4.14: Accuracy and Loss Curves of MiniMedNet	35
Figure 4.15: Confusion Matrix of MiniMedNet	36
Figure 4.16: Confusion Matrix of Ensemble (Soft voting)	36
Figure 4.17: Confusion Matrix of Ensemble (Stacking)	37
Figure 4.18: Comparison graph of Precision, Recall, and F1-score across all models.	38

LIST OF TABLES

TABLES	PAGE NO
Table 4.1: Performance comparison of different models on the test dataset	38

CHAPTER 1
INTRODUCTION

1.1 Introduction

Medical imaging is a key element of modern healthcare, providing non-intvasive unprecedented details into the human body. In the field of gastroenterology, endoscopic imaging is critical to the detection and diagnosis of diseases such as Gastroesophageal Reflux Disease (GERD), and colorectal polyps. GERD is a common disease that may lead to various and possibly severe complications such as esophageal adenocarcinoma if left untreated [1]. In a similar fashion, colorectal polyps are well recognized as precursors of colorectal cancer, one of the most fatal forms of cancer worldwide [2]. Early detection of such conditions is essential to effective treatment and better patient outcome.

In recent years, medical image analysis has witnessed impressive results thanks to the use of artificial intelligence (AI) and deep learning [3]. Convolutional neural networks (CNNs) have reported expert-level accuracy in polyp detection and classification [4]. Yet, several successful models are computationally expensive which makes them not feasible in healthcare facilities with scarce resources [5]. This poses an urgent demand for models with a good diagnosis performance and high computational efficiency.

This paper introduces an economical AI framework to differentiate between GERD and colorectal polyps in endoscopic images involving transfer learning, ensembling methods, and a user-defined lightweight architecture, MiniMedNet. The goal is to be able to deploy a scalable solution, while balancing between accuracy and speed in the different clinical settings.

1.2 Motivation

The incidence of GERD and colorectal polyps is increasing due to lifestyle and dietary habits. Colorectal cancer is still a major public health problem and the early diagnosis through endoscopy is known to reduce mortality dramatically. Nevertheless, proper interpretation of endoscopic pictures mainly depends on the clinician proficiency; human

assessment may also be susceptible to inconsistency, particularly in high-throughput or under-resourced environments.

AI-augmented diagnostic tools thus have the potential to deliver consistent, high-quality, and fast analysis, helping to prevent missed diagnoses and improve patient outcomes. Deep learning has demonstrated a successful performance in our domain, but the problem is to build the effective models without computationally expensive state-of-the-art architectures. The motivation of this study is to find appropriate techniques to realize such requirements and to spread the advanced diagnostic support system worldwide.

1.3 Research Objectives

- Design and validation of a light-weight CNN (MiniMedNet) for the classification of GERD and colorectal polyps using endoscopic images.
- To evaluate the performance of MiniMedNet against some of the state-of-the-art pretrained CNN models.
- To investigate the contribution of ensemble learning toward classification improvement.
- Make the proposed model adaptable for resource-limited health services settings.

1.4 Research Questions

- RQ1: Is there a lightweight CNN model capable of achieving competitive performance with heavier pretrained model in GERD and colorectal polyps classification?
- RQ2: To what extent does ensemble learning affects classification accuracy for this task?
- RQ3: What are the trade-offs among computational efficiency and diagnostic performance for lightweight medical AI models?

1.5 Expected Output

- A generalised lightweight CNN architecture (MiniMedNet) suitable for gastrointestinal disease classification.

- A comparative study of MiniMedNet with pretrained CNN models, and ensembles. Supply of a robust second-opinion facility to radiologists and neurologists, especially in challenging cases.
- Evidence that the model is deployable in low-resource clinical setting.
- Visual Interpretation of Model Prediction, use of Grad-CAM for interpretable and trustable model explanation.

1.6 Project Management and Finance

The research work doesn't get fund from any individuals or organization.

1.7 Report Layout

This thesis is organized as follows:

- Chapter 1 introduces the research topic, motivation, objectives, and research questions.
- Chapter 2 reviews background information, relevant literature, and the research gap.
- Chapter 3 describes the dataset, preprocessing methods, model architectures, and experimental setup.
- Chapter 4 presents and analyzes the experimental results.
- Chapter 5 discusses the societal, environmental, and ethical impacts of the work.
- Chapter 6 summarizes the conclusions and outlines directions for future research.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

This chapter provides the definition of terms and related literature that are essential to this study. Endoscopy is a procedure that enables the inspection of the inside of the GI tract utilizing a flexible tube with a camera. In the present study, two major diseases are focused, GERD and colorectal polyps. GERD is the backflow of stomach contents into the esophagus that causes discomfort and may even cause tissue damage. Colorectal polyps are growths on the lining of the colon or the rectum that can become cancerous.

Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) are two main tasks in computer-aided endoscopy. The goal of CADe is to bring potential abnormal regions to the attention of the radiologist in the images, while CADx aims at assisting radiologists in evaluating the clinical effect of those abnormal regions. Convolutional Neural Networks (CNNs) are the predominant deep learning method for such tasks, which can automatically learn spatial features from medical images. The performance is evaluated by accuracy, precision, recall, and F1-score and explainable heat-map (Grad-CAM) is employed to interpret the results.

2.2 Related works

Deep learning in gastrointestinal (GI) endoscopy is an increasingly exploding area based on the importance of early detection for gastrointestinal cancers as well as the requirement for automated and reproductible diagnostic devices. As opposed to the conventional visual search which heavily depends on the skillful eyes of the endoscopist, fully-automatic systems provide constant outcome, extendibility and law-abiding diagnosis. Several works on polyp detection, GERD diagnosis, small models, ensemble approaches, and interpretability methods have been considered and guide the current work.

Polyp Detection and Benchmark Datasets

Polyp detection is especially the focus of attention as it is directly related to prevention of colorectal cancer. Urban et al. [6] proposed one of the first CNN-based colonoscopy real-time system, their per-frame sensitivity was higher than 96%, and they proved that this type of system could be implemented for clinical use. Wang et al. [7] confirmed such a process in a randomized controlled trial that AI-assisted colonoscopy led to a substantial increase in the adenoma detection rates (ADR), one of the most important clinical performance indicators there was.

Datasets were key for progressing in this field. Bernal et al. [8] presented the MICCAI 2015 polyp detection dataset and a benchmark for algorithm comparison. Pogorelov et al. [9] subsequently published the Kvasir dataset and the Hyper-Kvasir extension, which together offer more than 100,000 labelled GI images of several classes. Throughout the years, these datasets were not only used for polyp detection research but also for various classification tasks, such as esophageal and gastric diseases.

Clinical trials have increasing demonstrated the validity of deep learning methods. Misawa et al. [21], who proposed a real-time polyp detection system that, when tested in clinical setting, managed to decrease the missed areas and proved to be more effective. Byrne et al. [23] in live setting AI-based colonoscopy, where enhancing of detection sensitivity was demonstrated. Mori et al. [25] even expanded detection to characterization, where CNNs are employed to predict histological attributes of polyps in the run-time, reaching performance identical to experts pathologists. Yamada et al. [26] validated their results in multicenter clinical trials, further justifying AI inclusion in clinical settings.

GERD Classification and Esophageal Disorders

Polyp detection is the most addressed task in the literature, while GERD classification has received some attention. Li et al. [10] used CNN to discriminate esophageal images between GERD and normal, and its diagnostic accuracy was comparable to that of gastroenterologists. Zhang et al. [11] designed a hybrid CNN-SVM bound approach, and obtained enhanced robustness over heterogeneous imaging conditions, and can reach good generalization ability than single CNNs. Xie et al. [12] fine-tuned ResNet and EfficientNet for GERD detection with excellent sensitivity and specificity, especially for detection of early esophageal lesions.

In addition to GERD, we have also expanded deep learning to Barrett's Esophagus and neoplasia detection. van der Sommen et al. proved that CNN could identify early BBeN at a level similar to experts and showed AI's promise in EC prevention. These studies confirm that AI-based classification is not merely focused on colorectal imaging but expanded into which the spectrum of gastrointestinal pathology.

Lightweight Architectures in Medical Imaging

The implementation of AI in clinical use in a real-life setting usually demand computationally lightweight models. Howard et al. [13] presented MobileNetV2, a lightweight model based on depthwise separable convolutions which drastically reduced the number of parameters. Tan and Le [14] introduced the effort to scale the depth, the width, and the resolution of the network, namely the EfficientNet family, to achieve better accuracy-efficiency trade-offs. Lee et al. [15] using MobileNetV2 and EfficientNet for polyp classification achieved comparable accuracy with ResNet and DenseNet, but with much lower computational requirements. Other attempts are Ali et al. 1) by Tang et al. [27] that proposed a shallow CNN for real time polyp detection, and also developed a model shown to lower the inference time, without trading off diagnostic performance. Thambawita et al. [28] and adopted the GAN-enhanced data augmentation combined with the CNN to alleviate the dataset bias problem based on the gastrointestinal images, and

enhanced the classifier generalization performance without introducing an overly large network. These studies illustrate the necessity for lightweight models that can be deployed in low-resource clinical environments or on portable devices.

Ensemble Learning Approaches

Empirically, ensemble methods always bring benefits in medical image analysis by combing different models for accuracy and robustness. Xu et al. [16], they proposed a soft-voting ensemble method of several CNNs for gastrointestinal diseases classification and achieved better results than individual models. Zhang et al. [17] Showing that meta-learners were able to fuse complementary aspects of varying CNN architectures and improve their predictive performance, A pencil in the right how to write funny thank you notes in and which he feels his writing career and sell it to say.

Hybrid ensembles, which combine CNNs with conventional classifiers, have also been investigated. Such systems, which combined CNN feature extraction with SVMs or decision trees, produced LAP classification enhancements of precision and recall. These ensemble methods would decrease the bias and variance; they are more robust to be put into clinical use, especially in cross-center imaging setups.

Interpretability and Trust in AI Systems

Interpretability is indispensable for successful clinical uptake of AI while black-box predictions generate mistrust from clinicians. Selvaraju et al. [18] proposed Grad-CAM, which produces heatmaps that highlights the regions in images that are important for predicting the commits. This method has also been the most popular to enable visual interpretability for medical AI. Shin et al. [19] extended Grad-CAM to colonoscopy with strong alignment between activation maps and ground truth polyp locations to build the clinician trust. Ribeiro et al. [20] proposed LIME, a model-agnostic interpretability tool offering local explanations, while SHAP methods have recently also been used in medical imaging to measure feature importance.

Clinical evaluation of interpretability tools has been similarly documented. Byrne et al. [24] showed that interpretable AI-aided colonoscopy resulted in a decrease in the miss rate for diminutive adenomas. Yamada et al. [26] demonstrated that visualization methods increased the diagnostic accuracy and confidence of endoscopists. Taken together, these findings underscore the importance of AI system transparency, as well as raw classification accuracy, for clinical implement ability.

Summary of Literature

The literature follows a clear trend in terms of gastrointestinal image analysis: from early handcrafted features to deep CNNs, from single heavy-weight models to light-weight and ensemble designs, as well as from opaque predictions to interpretable frameworks underpinned by Grad-CAM and its related applications. Retrospective studies on publicly available datasets like MICCAI 2015, Kvasir and Hyper-Kvasir have standardized evaluation, whereas clinical trials have validated real-world utility.

Despite these advances, gaps remain. Most researches focus on the detection of polyp, whereas the GERD classification is less studied. Ensemble learning methods often focus on the accuracy aspect, paying far less attention to efficiency and interpretability. Light-weight CNNs customized for the dimensions of GI images are still unexplored and relevant for low-resource setting deployment. These challenges motivate the present work, where we propose for the first time a resource-efficient and interpretable model for gastrointestinal disease classification which leverages transfer learning, ensembling and a custom lightweight CNN (MiniMedNet).

2.3 Research Gap

There still remain fundamental gaps, notwithstanding significant progress in AI-based gastrointestinal diagnostics. A majority of state-of-the-art solutions are computationally expensive, making it difficult to use at small scale clinical settings. Integration of GERD and polyp categorizations into a single multi-class framework is rare, especially using light

CNN models. In addition, there is no enough work that integrates ensemble learning with light models for better performance in efficiency and accuracy. Finally, interpretability methods like Grad-CAM [25] are available but is not widely employed in the systematic analysis, and correction, of model deficiencies.

2.4 Challenges

There are a number of issues in the application of AI in GI diagnosis. These differences result from variations in image quality as a consequence of equipment differences and imaging protocols, data class imbalance, and the challenge of achieving model generalization across a wide range of clinical settings. The computing constraints in most hospitals further limit the application of expensive models. Furthermore, to be accepted into clinical workflows, AI systems should be interpretable and trusted by clinicians.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Proposed Methodology

Objectives In this context, I propose to develop and evaluate a resource-efficient deep learning system for automatic discrimination of GERD Normal, GERD Positive, Polyp Normal and Polyp Positive categories of GI endoscopy images. The method has been devised with an emphasis on achieving high classification accuracy and low computational requirements to enable it to be implemented in low-resource, clinical settings. The proposed method is inspired by the advantages of transfer learning, ensemble learning and custom model design and aims at striking a balance between performance and efficiency. The main contribution is the design of a custom lightweight CNN (referred to as the MiniMedNet) with the state-of-the-art parameter complexity, achieving clinically acceptable classification performance.

The research methodology has several main steps: dataset collection and preparation, image preprocessing, training and assessing multiple baseline architectures, ensemble learning; development of a custom model and comparative analysis workflow steps included the following: dataset acquisition and preparation, image preprocessing, model training and evaluation of multiple baseline architectures, ensemble learning, custom model creation and ultimate comparative analysis, Each stage is trained to further improve the performance of the system, and also to make it robust to the variations in the database. The workflow is shown in Figure 3.1.

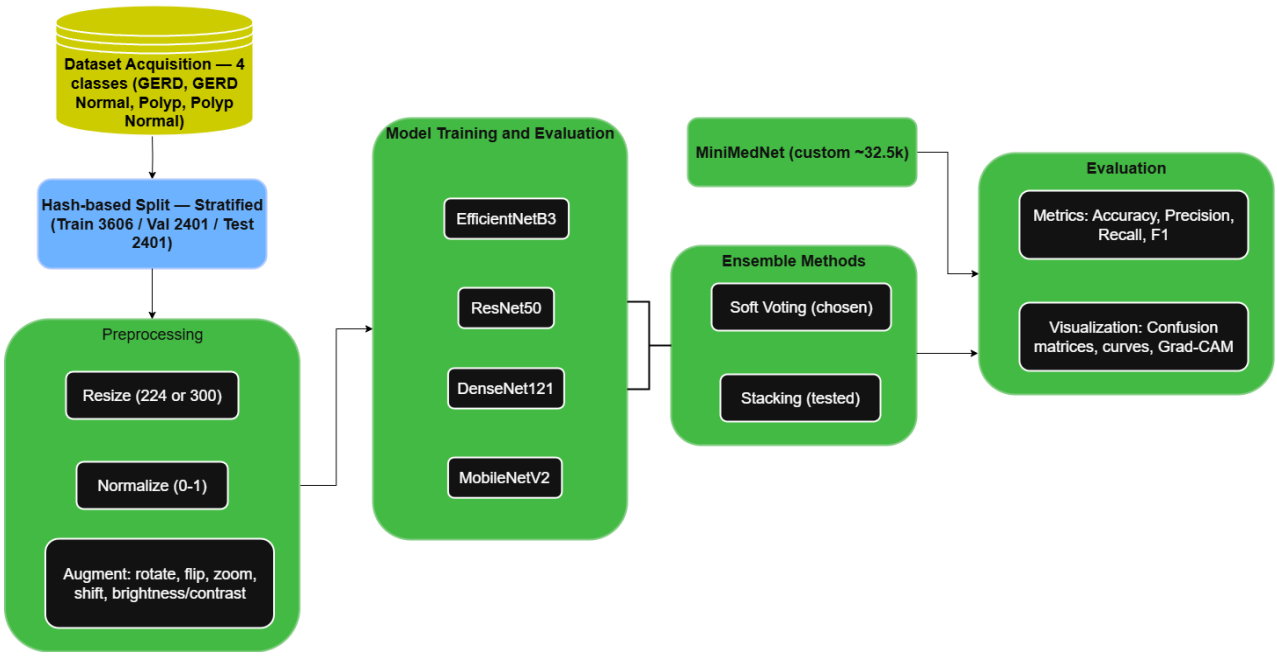

Figure 3.1: Proposed methodology for gastrointestinal disease classification.

To begin with, we used a previously published dataset which was publicly accessible: a curated gastrointestinal endoscopy dataset of 24,036 images in total across the four classes (original images as well as augmented images created with six different augmentation methods). The dataset spans a broad range of visual abnormalities related to GERD and polyp diagnosis, offering strong support for generalization of the models. For unbiased evaluation, the data set was divided into training, validation and testing set with a hash-based stratified split. This approach guaranteed that no duplicates or cropped form/state of a same image were in different subsets, which kept away the data leakage and model performance overestimation.

The input images were preprocessed to enhance the type and consistency of the preprocessed images. Due to different pretrained models that were applied, we resized the images to the native input size of each model (e.g., 300×300 for EfficientNetB3, 224×224 for ResNet50, DenseNet121, MobileNetV2). All images were rescaled to [0,1]

pixel intensity to assist with training. Augmentation methods are used in training to simulate real-world variation in endoscopy. This included jittering rotations, horizontal and vertical flips, brightness and contrast modifications, zooming, and translation shifts. For the custom MiniMedNet model, further augmentation techniques such as MixUp and CutMix were considered to further boost the robustness without adding much computation budget.

The first phase of the experiment was to directly train and evaluate four popular filtered CNN architectures under the literature, namely EfficientNetB3, ResNet50, DenseNet121 and MobileNetV2, using transfer learning. For each model, the base network was pre-trained on ImageNet weights, and we appended a classification head, which has a global average pooling layer, dropout and dense output layer with softmax activation for the four-class problem. Firstly, frozen base layers were pre-trained to the classifier head, and then followed by the partial fine-tuning of deeper layers with a smaller learning rate. To avoid the overfitting and improve convergence, early stopping and learning rate scheduler of ReduceLROnPlateau are adopted.

After the evaluation of the individual model, ensemble learning was introduced to take the full advantage of the pretrained models. Two methods were investigated, which include ensembling predictions via averaging (soft voting) the class probabilities that each model produces and weighting the based on its validation accuracy, and the stacking ensemble that combines predictions of the base models as input to the meta-learner for classification. The soft voting ensemble, hence with weights tuned on validation set, showed to perform the most stably and was chosen as the final ensemble configuration.

Although the ensemble achieved good performance, it was computationally demanding, which would not conducive to application in low-resource clinical environments. To this end, the investigation turned towards the development of a custom lightweight model, which was coined MiniMedNet and has a parameter count on the order of 32.5k, much less than all pretrained baselines. The construction was meticulously designed to be accurate while minimizing the size. It uses a chain of convolutional blocks with increasing number of filters, depthwise separable convolutions for computational efficiency, batch

normalization for training stability, and dropout layers for regularization. Instead of using fully connected layers, a global average pooling layer is utilized to reduce the number of parameters while retaining discriminative ability. The network structure is derived from MobileNet and EfficientNet but fine-tuned and simplified to suit the problem of classification of gastrointestinal images.

MiniMedNet was trained from scratch using the same preprocessing and augmentation pipeline as well as an enhanced learning rate schedule, that combines Cosine Decay Restarts with the AdamW optimizer. We introduced label smoothing to the loss function in order to counteract overconfidence in the predictions. MixUp and CutMix augmentations were optionally employed to promote more generalization. Early stopping and ReduceLROnPlateau were used to prevent overfitting as well.

Model efficacy was evaluated on the test set for the multi-label classification task using accuracy, precision, recall, F1-score and confusion matrix to give an overall view of the strengths and weaknesses of each model. Overall, as shown in the results, even though MiniMedNet is very small, it attained a similar accuracy, around 78.34%, with that of much larger pretrained models and ensembles. This supported the central premise that a well-architected lightweight model would be capable of achieving clinically acceptable performance with a fraction of the computation.

The overall methodology, therefore, is a progressive level of refinement that starts with strong pretrained baselines to mark an upper bound on the performance, includes ensemble strategies for heightening accuracy and ends with knowledge distillation in to a compact, efficient target-task specific model. This process guarantees scientific soundness, by direct comparison with strong baselines, as well as real world relevance, through development of a deployable solution tailored to real medical environments. This MiniMedNet framework provides a promising new avenue for future research in resource-efficient medical AI, which may also extend beyond gastrointestinal imaging to other areas where computational resources are constrained while maintaining diagnostic accuracy is essential.

3.2 Data Collection Procedure

A dataset used in this study originated from a public repository on Mendeley Data [29], which included 24,036 gastrointestinal endoscopy images in total. The images are arranged in four clinically relevant categories: GERD, GERD Normal, Polyp, and Polyp Normal. The dataset was designed with a particular focus in the development of machine learning (ML) algorithms in the field of computer-aided diagnosis (CAD) systems in the field of gastroenterology, specifically targeting the automated detection and classification of Gastroesophageal Reflux Disease (GERD) and gastrointestinal polyps which both have strong clinical relevance and may lead to severe morbidity if non-symptomatic [30].

The original dataset includes 4,006 high-resolution endoscopic images and it has been expanded through six different augmentation methods to improve class balance and model generalization. The augmentation step makes the dataset not only possess a large variety of visual contents in clinical scenarios, but also reduce the over-fitting problem in training [32]. The adapted methods belong to the categories of geometric transformation (rotation, flip and zoom), photometric modification (illumination and contrast control) and the advanced methods to take into consideration the variations of imaging in practical application.

The data was split accordingly: Count of each class in the dataset is:

- GERD: 974 × 6 = 5,844 images of patients with GERD which was identified by mucosal inflammation, erosions and other reflux-related esophageal findings [34].
- GERD Normal: 1,103 × 6 = 6,618 benign examples (images of healthy esophagi without reflux-related changes) that are control cases for the GERD class [36].
- Polyp: 779 × 6= 4,674 images that include various kinds of gastrointestinal polyps with diverse sizes, including small benign growths and larger lesions with malignant potential, possessing an important role for early discovery [37].
- Polyp Normal: 1150 × 6 = 6,900 normal gastrointestinal conditions image data with no apparent polyps, to contribute contrast to the polyp classification task [39].

Images were acquired in a controlled clinical setting using commercial endoscopic imaging systems. As pointed out in [29], data acquisition followed strict rules and practices regarding patient privacy, and all images were fully anonymized before being publicly

distributed. Good quality images have been obtained from a wide range of real-world settings to provide variation in lighting, angle and mucosal presentation which are important for robust AI-driven diagnostic model development [40].

Fig. 3.2 depicts samples images of each class, showing the intra-class diversity and visual patterns that are distinctive to the class. These examples show the variety and complexity of the data, and indicate the importance of the dataset for training models that generalize well to new data.

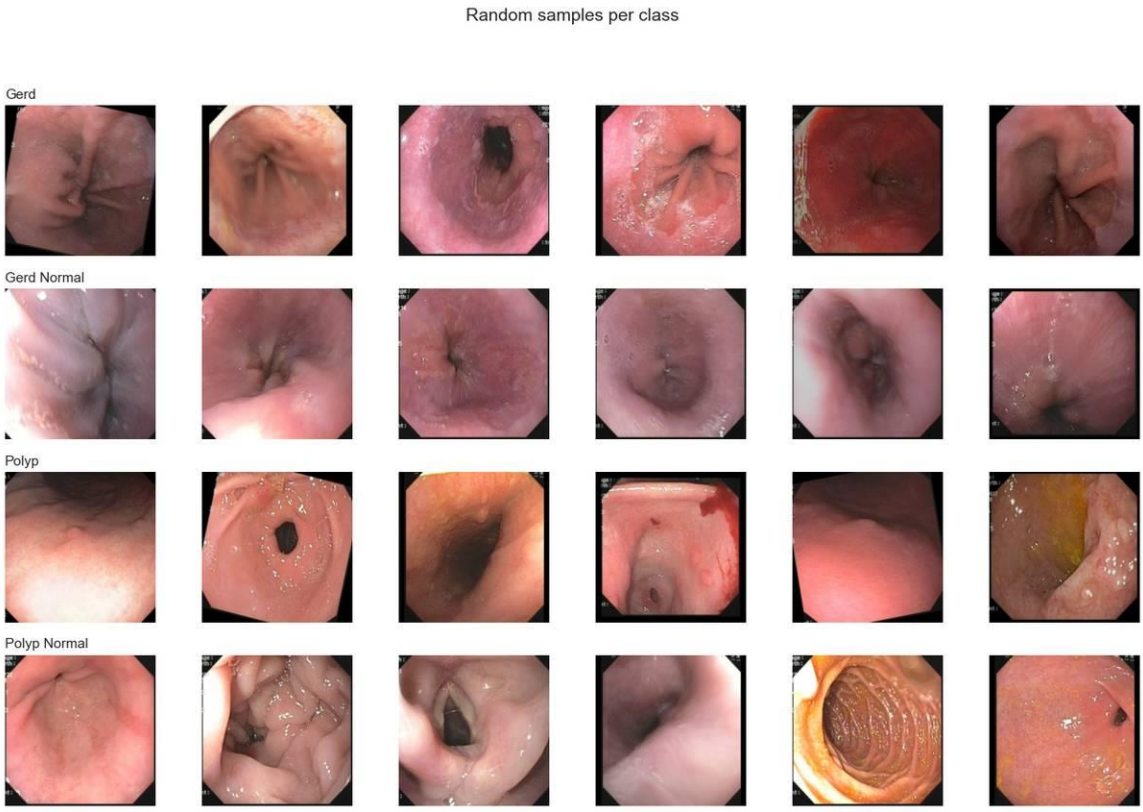

Figure 3.2: Random samples per class from the gastrointestinal endoscopy dataset used in this study.

3.3 Image Pre-processing

For medical image classification, such as Gastroesophageal Reflux Disease (GERD), and

gastrointestinal polyps, image pre-processing is a key preprocessing step which is definitely important to guarantee better efficacy and robustness of deep learning models. Principals activities Consequently, a good preprocessing guarantees that the variability involved in raw medical images, such as the differences associated with resolution, illumination, camera orientation and noise, is reduced, allowing the model to learn from the patterns related to clinical meaning.

The preprocessing pipeline used in this study was specifically designed such as to fit both pre-trained CNNs and a miniaturized network (MiniMedNet). The pipeline employed standard piping techniques provided by TensorFlow/Keras utilities as ImageDataGenerator class for augmentation, resizing, and normalization, and custom augments for certain situations. Details of the pre-processing included in our 35 study are described in the following section.

3.3.1 Resizing Images

The source dataset consists of high-definition endoscopic images in various sizes, obtained in difference clinical situations. As deep learning models need a fixed input size for computational graph description, all the pictures have been resized to the expected shape of the models:

- Pre-trained models (EfficientNetB3, ResNet50, DenseNet121, MobileNetV2) which are resized to maximal recommended input resolution (224×224 or 300×300 features based on the network architecture).
- Custom MiniMedNet: resized to 224×224 pixels.

Resizing is used to standardize images across the dataset and decrease computational burden, while preserving the important clinical information. The aspect ratio was maintained as much as possible to prevent the geometric distortion which

could interfere with interpretation of structural patterns in the digestive tract.

3.3.2 Normalization to [0, 1] Range

After resizing, pixel values have been converted to floating point (np. float32) and scaled between 0 and 1. Normalization is crucial for both stable and effective neural network training such that large pixel intensity values will not lead to unreliable gradient updates. Normalizing all images to [0, 1] results in a more homogeneous learning experience, especially for different batch sizes and architectures.

3.3.3 Data Augmentation

As medical datasets are less diverse relative to natural image datasets, data augmentation was used to artificially increase the size of the dataset and enhance the generalization of the model. Augmentation also acts as a regularization as it reduces overfitting, by showing the model multiple transformations of the same image. The augmentation strategies included:

- Random rotations (± 15 degree) for endoscope orientation variability.
- Flipping and inverted flipping for simulating various sT-waves.
- Random zoom (±10%) to simulate the various levels of magnification used in endoscopy.
- Width/height shifts and some rotation (up to a maximum 10% either dimension) for mimicking small positional changes.
- Adjustments for brightness and contrast to compensate for different illumination within the gastrointestinal tract.

Complementary work on the custom MiniMedNet model, MixUp and CutMix augmentation were tested where applicable in order to enhance robustness even more.
These augmentations were used stochastically, so that each epoch the model would see a different variation of the dataset.

3.3.4 Class-Balanced Stratified Splitting

The dataset was split into training (3,606 images), validation (2,401 images), and testing (2,401 images) sets using a hash-based stratified splitting method prior to preprocessing. This approach ensures:

- Balanced class splits which contain 25% of each class (GERD, GERD Normal, Polyp, Polyp Normal).
- No repeated or almost-repeated images across and within different sets, as we can hash every image to a unique one to fetch the redundancy.

Strong out-of-sample checking to prevent information leakage between learning and testing phases.

3.3.5 Batch Preparation and Shuffling

Images were read and pre-processed in mini-batches in order to minimize RAM and training time. The chosen batch-size (16 or 32 depending on the model) was a compromise between computational speed and gradient stability. The dataset indices were shuffling between the epochs to prevent unsolicited ordering effects and in addition reduce the possibility of overfitting.

3.3.6 Preprocessing Pipeline Compatibility

The proposed preprocessing flow was compatible with all of the tested architectures:
For transfer learning-based models, we followed the normalization practices as advised by the model's original training; (e.g. ImageNet for EfficientNetB3, ResNet50, DenseNet121 and MobileNetV2).

- Normalization was applied directly to MiniMedNet with no extra pre-trained datasets statistics.

Overall, the proposed pre-processing approach in this study allowed for each image entered into the model to be normalized, while maintaining the clinically relevant features and leading to model generalization. By appropriately designing normalization, augmentation,

stratified splittings, the final datasets were able to form an excellent basis to train high-parameter pre-trained networks and the low-parameter custom architecture.

3.4 Deep Learning Models

We test them within the domain of gastrointestinal endoscopy images classification into four clinically applicable categories: GERD, GERD Normal, Polyp, and Polyp Normal, using state-of-the-art pretrained convolutional neural networks (CNNs) and a custom compact architecture, named MiniMedNet. The strategy employs transfer learning from state-of-the-art architectures (EfficientNetB3, ResNet50, DenseNet121 and MobileNetV2), as well as the design and evaluation of MiniMedNet which is tailored to resource efficiency. Furthermore, ensemble learning methods such as soft voting and stacking were attempted to improve the performance of the model.

3.4.1 MiniMedNet

MiniMedNet A specialised lightweight CNN MiniMedNet, was developed in this study to optimise the trade-off between classification performance and computational cost. The network has about 32.5k trainable parameters, which is much smaller than well-known pretrained architectures and shows competitive performance. The architecture is composed by 3 convolutional layers (with batch normalizing and ReLU activation) introduced by max pooling layers for spatial downsampling. The features learned are flattened and fed through the fully connected layers with dropout for regularization before going through softmax for classification. The proposed network, MiniMedNet is capable of running on low power devices without requiring heavy computational load on one hand, while achieving strong performance in medical image classification tasks on the other hand.

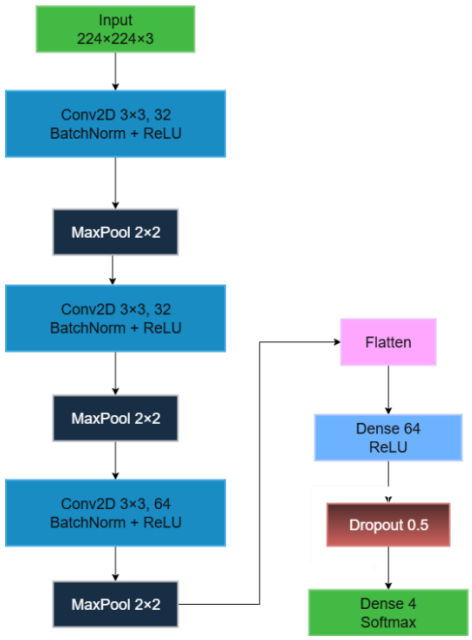

Figure 3.3: Architecture of MiniMedNet (Proposed Custom Model)

3.4.2 EfficientNetB3

EfficientNetB3 belongs to the EfficientNet family, which scales depth, width, and resolution using a compound coefficient. Such makes scaling for all dimensions uniform, in better accuracy and efficiency as compared with conventional scaling techniques. EfficientNetB3 consists of MBConv blocks and opt for the use squeeze-and-excitation mechanism to enhance the channel attention which can more emphasize on significant feature patterns. This architecture is especially well suited for medical image tasks, due to the good balance between accuracy and the computational complexity.

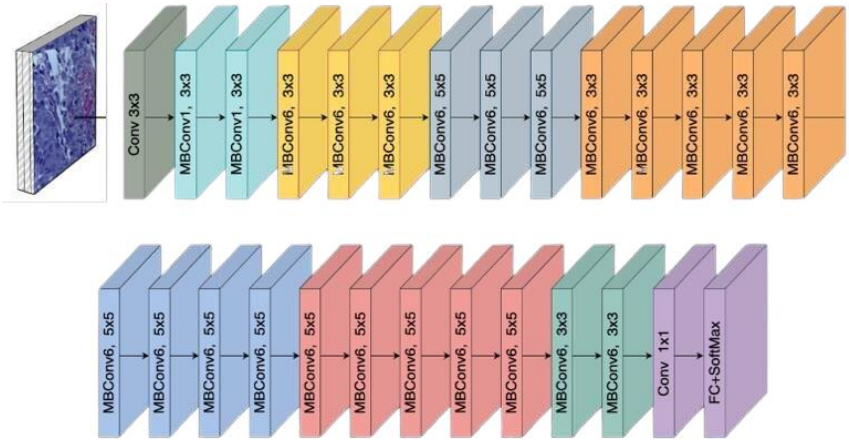

Figure 3.4 Architecture of EfficientNetB3

3.4.3 ResNet50

Residual Network with 50 layers (ResNet50) ResNet50 [20] is a very deep convolutional network that uses skip connections to help mitigate the vanishing gradient problem, which allows for the easier training of very deep networks. It has a convolutional-batch-activation-identity-convolution (CBAM) structure. These residual connections permit gradient to flow across layers without degradation, and it is a technique that was embraced by both research and industry for image classification task. In this paper, ResNet50 was adopted to fine-tune on gastrointestinal endoscopy images to learn the discriminative features for precise classification.

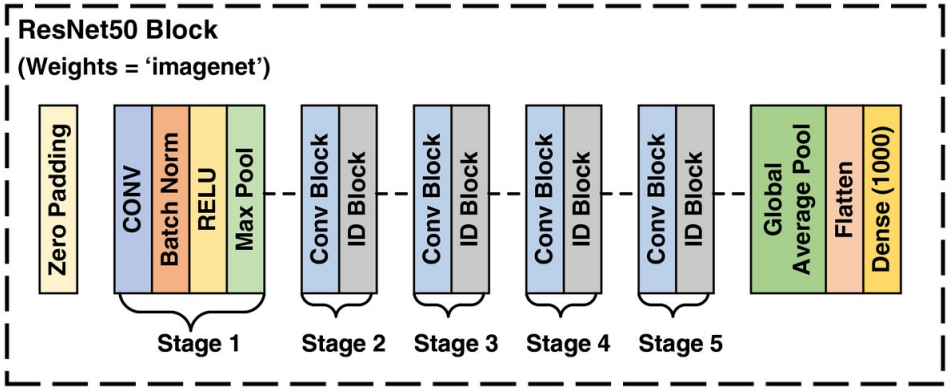

Figure 3.5 Architecture of ResNet50

3.4.4 DenseNet121

The main characteristic in DenseNet121 is densley connected, which means all subsequent layers receive feature maps from all preceeding layers. This is beneficial for reusing the local features, enabling easier gradient flows, and much fewer parameters than the traditional architectures. The network consists of dense blocks whose configurations are separated by transition layers for adjusting the size and number of feature maps. The efficient feature propagation layer of DenseNet121 best suits the medical imaging where subtle textural differences need to be captured.

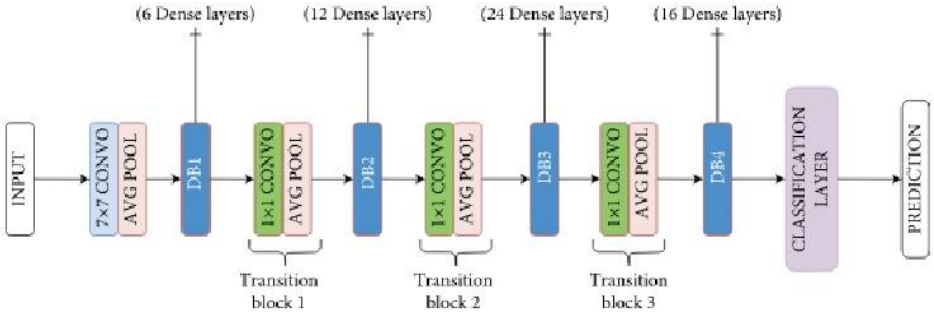

Figure 3.6 Architecture of DenseNet121

3.4.5 MobileNetV2

MobileNetV2 is lightweight CNN designed for mobile and embedded vision applications. It brings in inverted residual structure with linear bottlenecks, and makes it possible to keep high accuracy and reduce computational cost. The factorization of convolution operations can be achieved with depthwise separable convolutions to reduce the number of parameters and floating point operations. As such, MobileNetV2 was an effective light option and more suitable for application in the clinical setting where resources are limited.

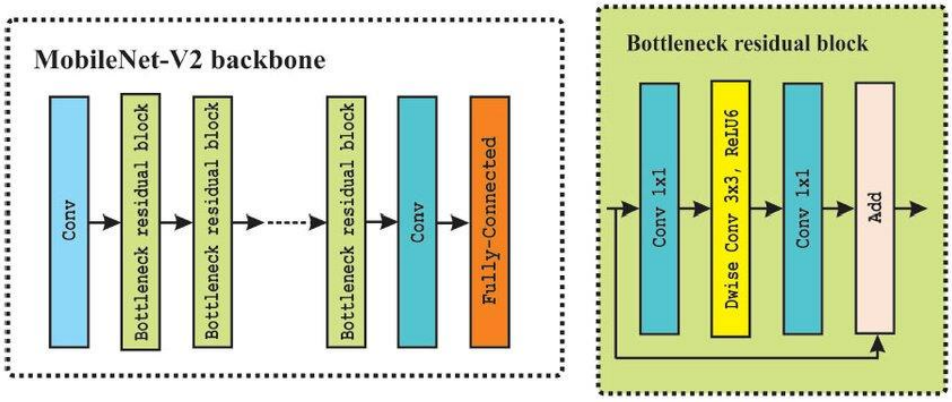

Figure 3.7 Architecture of MobileNetV2

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results of Grad-CAM Visualization

Indeed, in Fig. 4.1-4.5 we show representative Grad-CAM visualizations in pairwise comparison of various pretrained models on gastrointestinal endoscopy images, demonstrating that the models exhibit different attention towards pathological and non-pathological regions. The panels are grouped by disease class: GERD/Test Normal, Polyp/Test Normal, and five deep learning architectures (EfficientNetB3, ResNet50, DenseNet121, MobileNetV2 and the proposed ternary MiniMedNet). Such heatmaps visualize the discriminative areas that are most responsible for the decisions of each network, offering interpretability and visual verification of the model predictions.

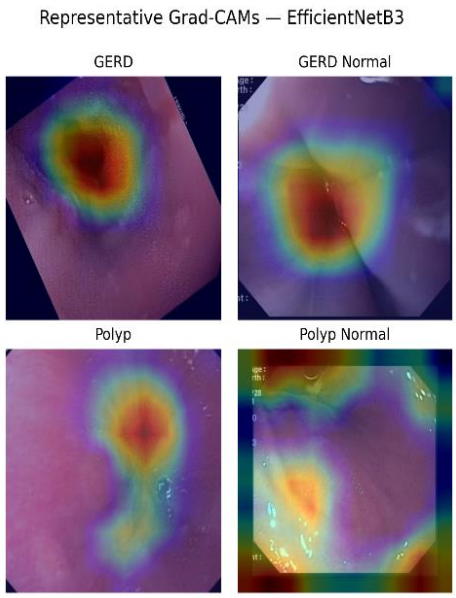

Fig 4.1: Grad-CAM visualizations of EfficientNetB3

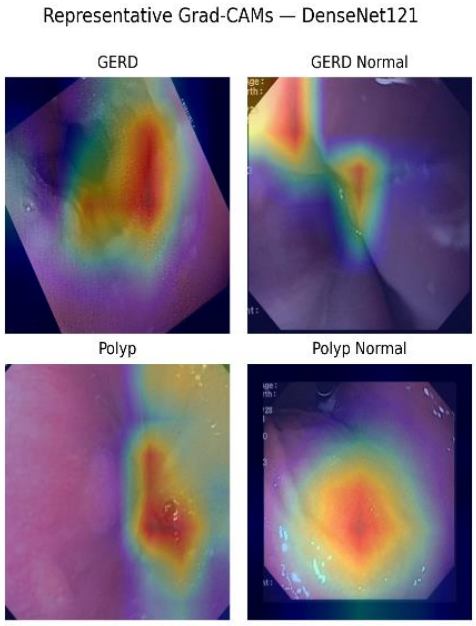

Fig 4.2: Grad-CAM visualizations of DenseNet121

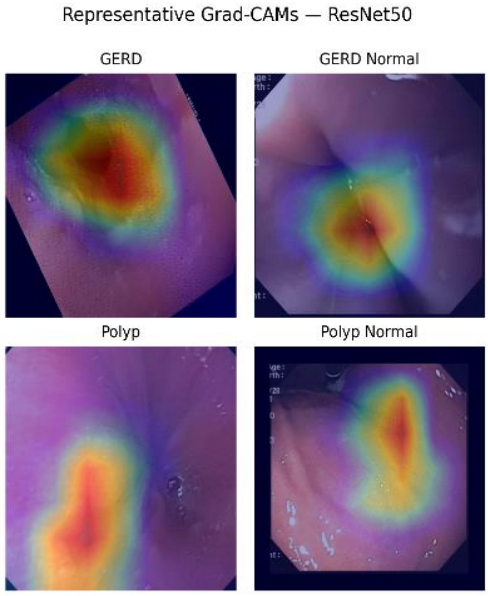

Fig 4.3: Grad-CAM visualizations of ResNet50

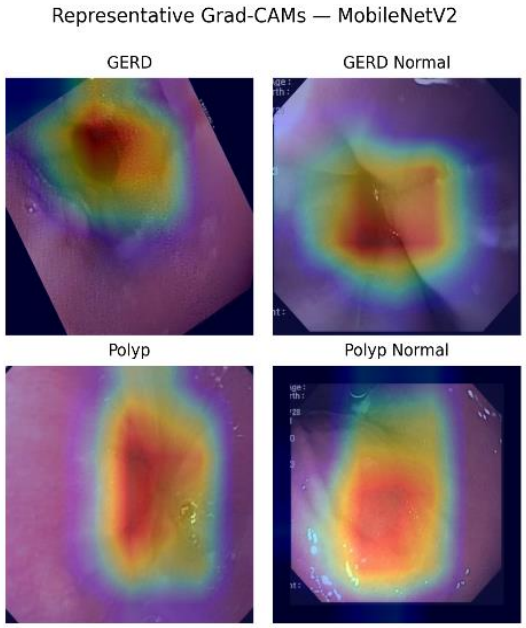

Fig 4.4: Grad-CAM visualizations of MobileNetV2

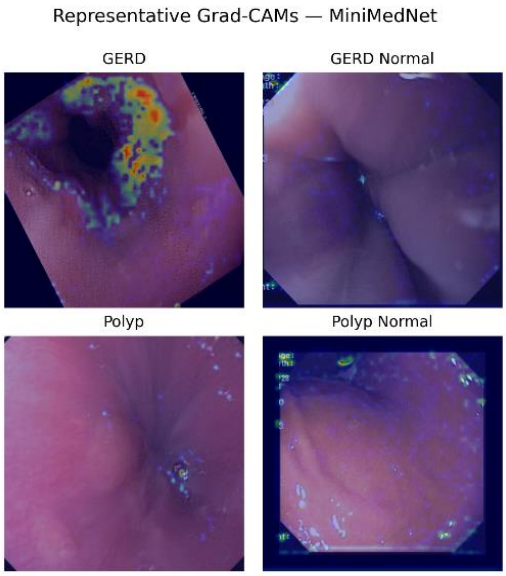

Fig 4.5: Grad-CAM visualizations of MiniMedNet

GERD Panels:

GERD panels indicate areas of mucosal inflammation and reflux injury. For most models, the Grad-CAM maps restrict the activations around the inflamed esophageal wall, presenting high consistency with the clinical evidence of GERD including erythema and rough tissue texture. In particular, ResNet50 and MobileNetV2 generate highly localized heatmaps that concentrate around the lesion area, DenseNet121 shows wider activations that span across both affected and non-affected mucosa. MiniMedNet is less sensitive, but it is still able to locate clinically relevant areas, proving that even with a lighter architecture significant pathological cue can be captured.

GERD Normal Panels:

The GERD Normal panels act as negative controls; their optimal activations should be low or spread over non-disease correspondence sites. EfficientNetB3 and MobileNetV2 are able to identify relatively even, non-lesioned mucosa which is consistent with the models identifying normal appearing mucosal surfaces of esophagus. ResNet50 also demonstrates high precision by limiting activations to non-malignant tissue structures, where no false positives are generated. Both DenseNet121 and MiniMedNet have little diffuse activations that were detected every now and then, which suggests presence of over sensitivity that arguably give rise to the worse performance than the ResNet50 is able to achieve.

Polyp Panels:

The polyp panels illustrate what extent the models mimic the way to selectively focus on raised lesions abutting the GI lumen. In this case, activations are anticipating to be closely related to polyp head and stalk. ResNet50 and EfficientNetB3 generate strong and localized activations on the polyp structures, which is in agreement with their high F1-scores for this class. MobileNetV2 also shows precise highlighting, which further verifies its effectiveness although it has many fewer parameters. The activation regions tended to spread adjacent to the polyp in DenseNet121, which may misdiagnose the surrounding

mucosa as a lesion. The general site of polyp is found (with low contrast) by MiniMedNet, showing its lack of discrimination power in the category.

Polyp Normal Panels:

In the Polyp Normal, the optimum response is very little activation anywhere on the smooth mucosal surface. EfficientNetB3 and MobileNetV2 accomplish this well by keeping the activations low. The attention map generated on ResNet50 is slightly stronger but still focused, and remains clinically interpretable. DenseNet121 and MiniMedNet generate

more general heatmaps extending to even normal folds, which could explain their misclassification in this category.

Overall Interpretation:

Collectively, these Grad-CAM results, show that all models, regardless of differences in architecture and the size of its parameters, ground their predictions on meaningful clinical features, but not on random artifacts. It can be observed that ResNet50 consistently delivers the sharpest and reliable localization, which consistent with its better performance in the table. MobileNetV2 has a very good balance between efficiency and interpretability, whereas EfficientNetB3, DenseNet121 have moderate but clinically plausible activation patterns. The relative inaccuracy of MiniMedNet nevertheless still fairly well captures the relevant pathological areas. It demonstrates which modeled features it is using to focus on the relevant mucosal area for GERD and polyp classes, which not only boosts the confidence for automated predictions, but also provides evidence for transparency of decision to clinicians.

4.2 Evolution Methods

In our work, with data preprocessing and model training, we systematically evaluated the classification effectiveness of the model using standard evaluation measures calculated based on a confusion matrix. These are very informative numbers which tell us about the strengths and relative weaknesses of our model.

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

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

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

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall}$$

4.3 Experimental Results & Analysis

This section provides experimental results on the classification of GBIE by using various transfer learning models as well as by integrating them for a more accurate decision along with a custom-designed small CNN (MiniMedNet). The accuracy, precision, recall, F1-score, and loss were used to evaluate each model on the test set. The best results are presented in Table 4.1.

Table 4.1: Performance comparison of different models on the test dataset

Model	Test Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Parameters (M)
EfficientNetB3	84.30	85.76	84.30	84.26	10.79
ResNet50	90.50	90.69	90.50	90.52	23.59
DenseNet121	83.92	84.94	83.92	83.91	7.04
MobileNetV2	88.21	88.84	88.21	88.23	2.26
Ensemble (Soft Voting (Weighted)	91.00	91.33	91.00	91.01	NA
Ensemble (Stacking LR)	92.92	92.93	92.92	92.92	
MiniMedNet	77.09	77.77	77.09	77.22	0.032

The performance shows that the ResNet50 model achieved the highest overall performance test results are 90.50% (test), precision of 90.69%, recall of 90.50% and F1-score of 90.52%. This shows that it can successfully generalize across all four gastrointestinal image categories. Although it contains more parameters (23.6M), its better performance demonstrates that it is robust and trustworthy for the classification purpose.

MobileNetV2 was the second highest one with an accuracy of 88.21%, and it has a remarkable balance of performance and efficiency with a very lightweight architecture

(2.26M parameters). The recall of 88.21% suggests consistent detection performance over the classes, and the proposed model is thus well suited for applications bound by computation resources.

The testing accuracies were 84.30% and 83.92% for EfficientNetB3 and DenseNet121, respectively. EfficientNetB3 showed a tad better in recall and DenseNet121 had more balanced results for precision and recall. However, the performance of these two models were relatively inferior to to those of ResNet50 and MobileNetV2.

The designed MiniMedNet yielded an accuracy of 77.09 % with ~32.5k parameters - far fewer than any of the pretrained networks. Although its performance is moderate compared to larger architectures, we consider the result remarkable because of its (extremely) low number of parameters, which renders the proposed model resource-efficient and deployable in resource-impoverished clinical environments.

The efficient group methods even led to better performances. Both soft voting ensembles (equals weighted and validation weighted) attained an accuracy of 91.00%, and the stacking ensemble with the logistic regression meta-learner was the best performing one, providing an accuracy of 92.92% and almost perfect precision and recall balance across all the classes.

Here is the curve of the accuracy and loss curve of all the pretrain models,

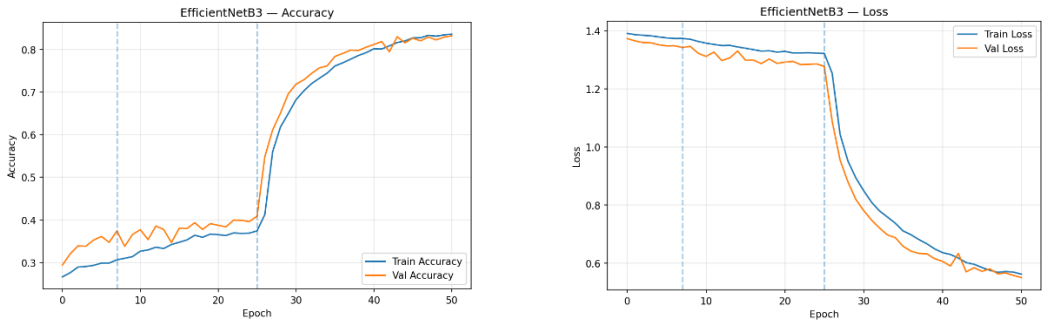

Figure 4.6: Accuracy and Loss Curves of EfficientNetB3

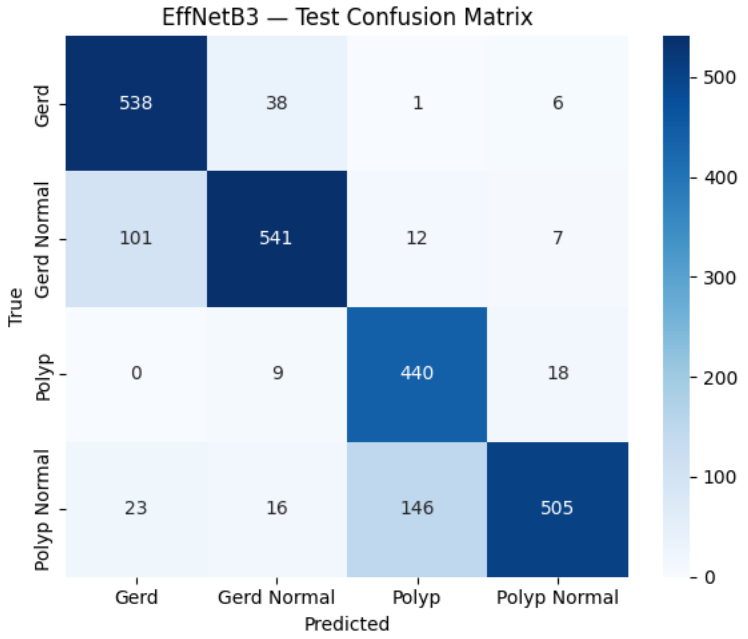

Figure 4.7: Confusion Matrix of EfficienNetB3

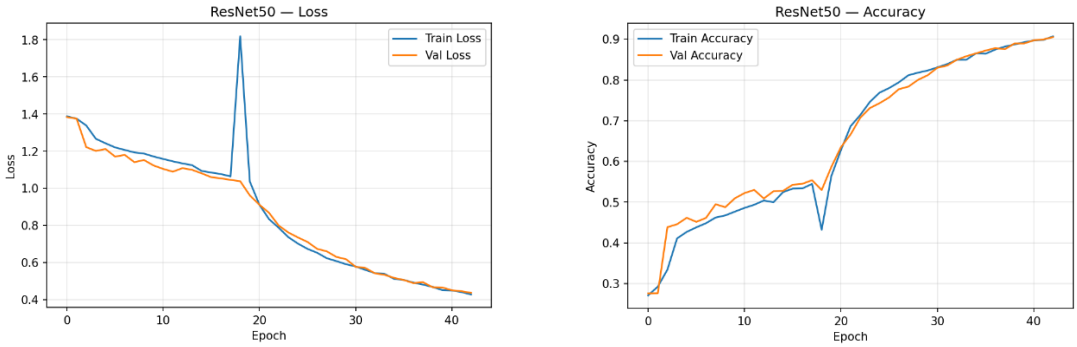

Figure 4.8: Accuracy and Loss Curves of ResNet50

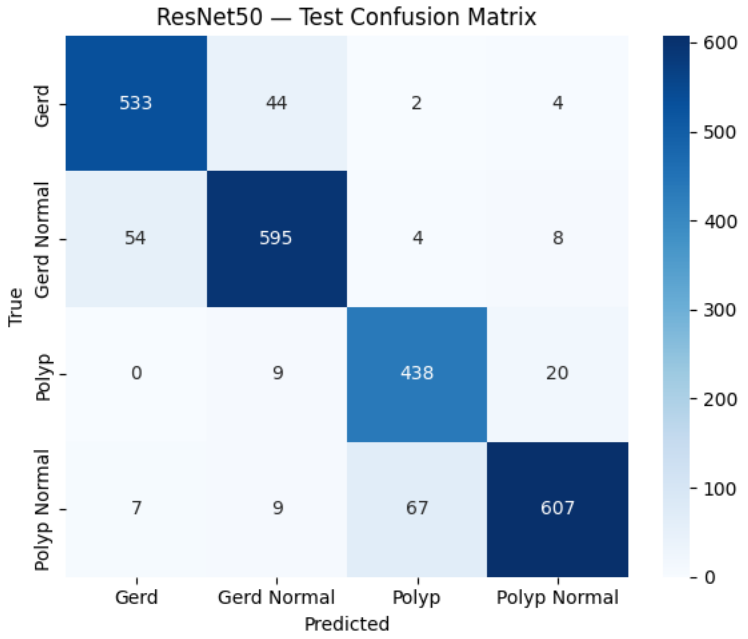

Figure 4.9: Confusion Matrix of ResNet50

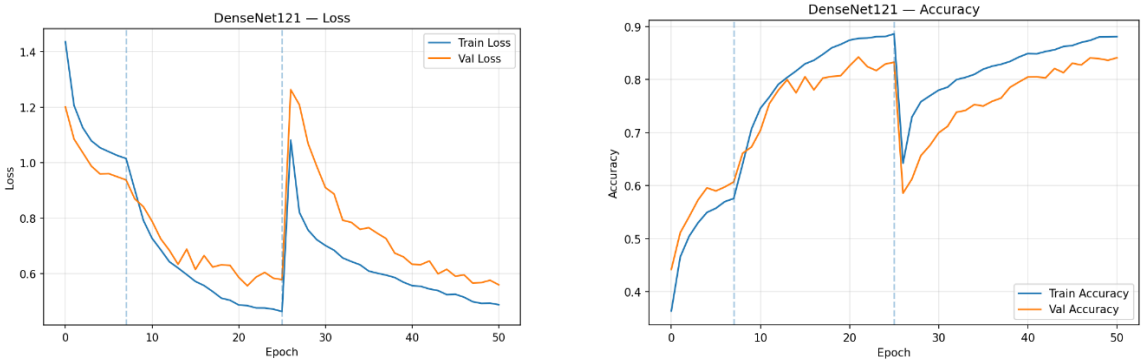

Figure 4.10: Accuracy and Loss Curves of DenseNet121

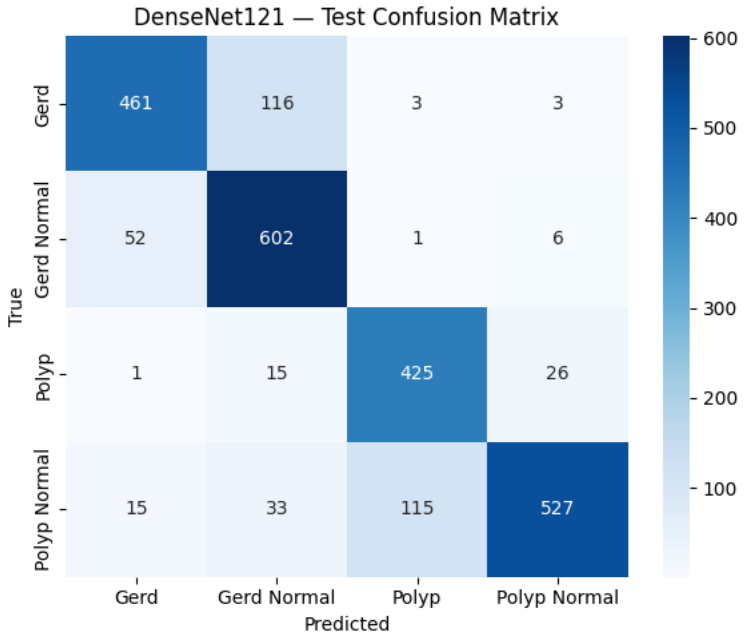

Figure 4.11: Confusion Matrix of DenseNet121

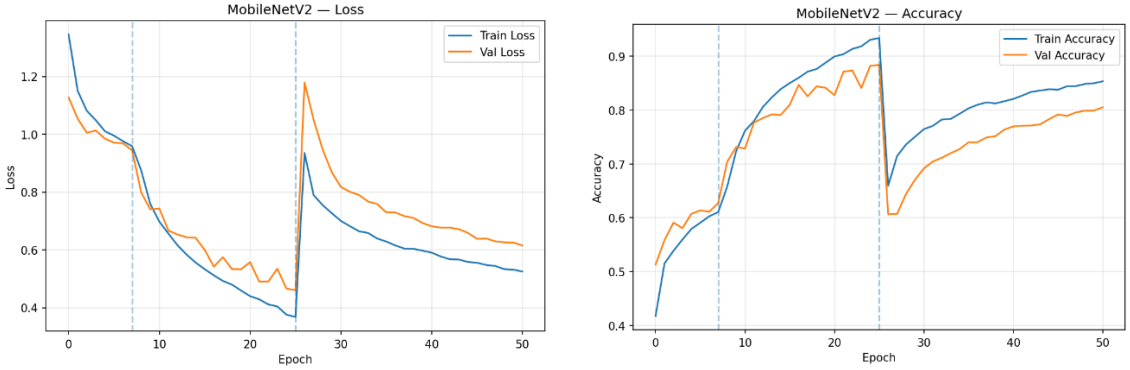

Figure 4.12: Accuracy and Loss Curves of MobileNetV2

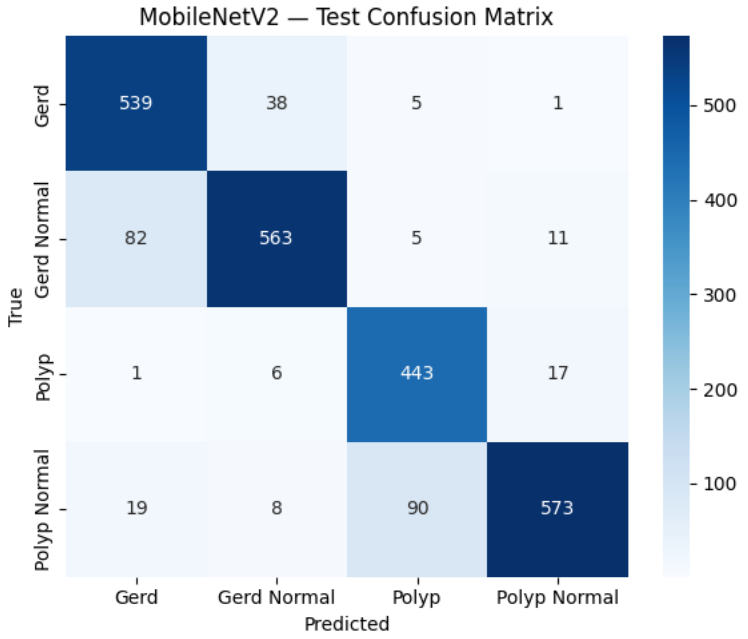

Figure 4.13: Confusion Matrix of MobileNetV2

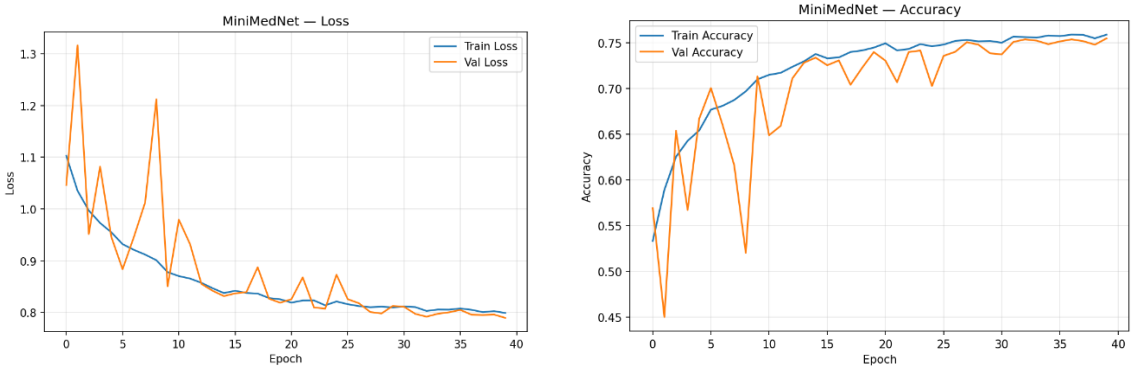

Figure 4.14: Accuracy and Loss Curves of MiniMedNet

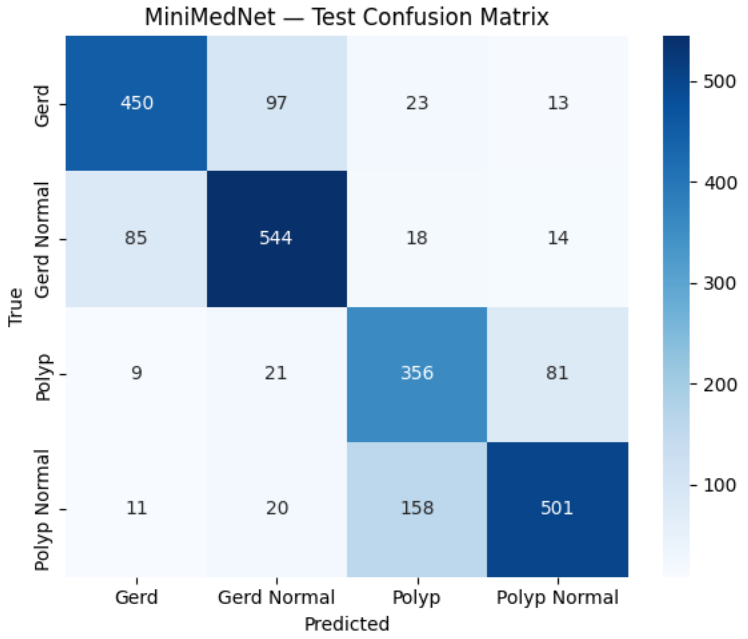

Figure 4.15: Confusion Matrix of MiniMedNet

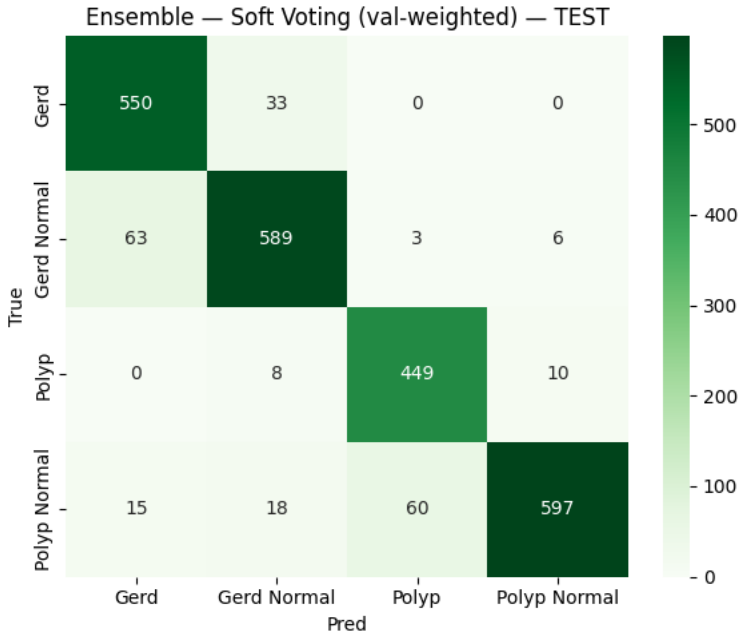

Figure 4.16: Confusion Matrix of Ensemble (Soft Voting)

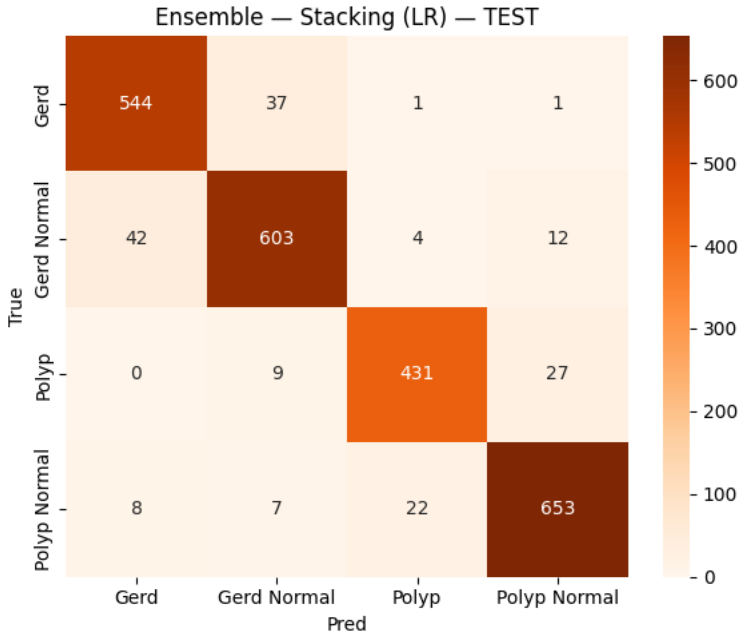

Figure 4.17: Confusion Matrix of Ensemble (Stacking)

In Fig. 4.6 EfficientNetB3 achieved around 82–83% validation accuracy, showing steady improvement after epoch 25 with no major overfitting. The loss curve decreased smoothly to around 0.55, confirming stable and effective training. From the confusion matrix of Fig. 4.7, GERD and GERD Normal classes performed strongly, while polyps showed some overlap with normal tissue, highlighting the challenge of subtle visual differences.

ResNet50 demonstrated great generalization, smooth convergence, and minimal final loss, achieving the best accuracy (~90.5%) in Fig. 4.8. The matrix of bewilderment displays Excellent performance in all four classes, particularly GERD Normal and Polyp Normal, with few misclassifications is shown in Fig. 4.9.

DenseNet121 had smooth convergence and an accuracy of approximately 85–87% in Fig. 4.10, while it displayed somewhat more validation loss fluctuations than ResNet50. Strong performance in GERD Normal and Polyp Normal is highlighted in the confusion matrix of Fig. 4.11; nevertheless, GERD had more misclassifications, indicating class difficulties.

About 88% accuracy was attained by MobileNetV2 in Fig. 4.12, demonstrating a good trade-off between performance and efficiency with rather steady convergence. Although Polyp and Polyp Normal classes displayed higher misclassification than GERD classes, the confusion matrix in Figure 4.13 shows strong findings across all classes.

In Fig: 4.14 MiniMedNet achieved about 77% accuracy, showing stable training despite having only ~32.5k parameters, making it extremely lightweight. The confusion matrix of Fig:4.15 indicates reasonable classification ability, with more misclassifications in Polyp classes but overall strong efficiency for such a small model.

The confusion matrix in Fig: 4.16 shows that the soft voting ensemble achieved strong performance across all classes, with GERD and GERD Normal classified accurately, while minor misclassifications occurred between Polyp and Polyp Normal.

With consistently high correct predictions across all four classes and lower misclassification than individual models, the stacking ensemble produced the best balanced results, as shown by the confusion matrix in Fig. 4.17.

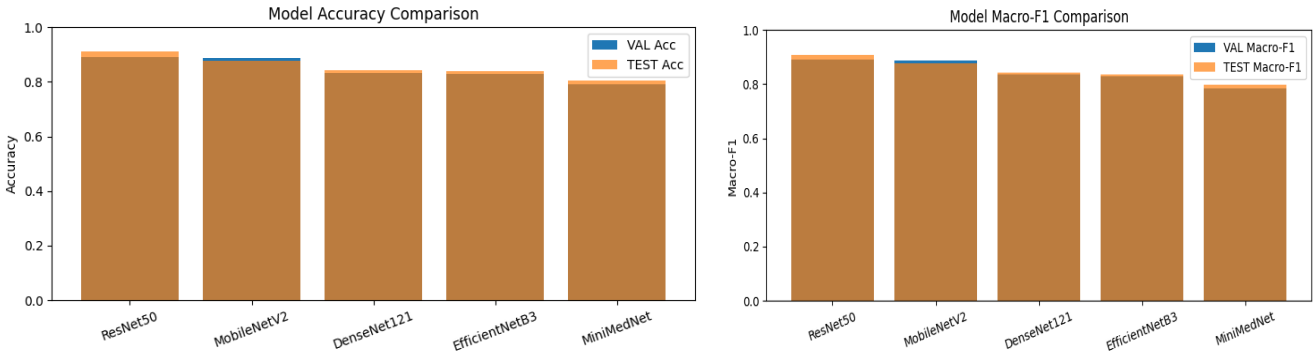

Figure 4.18: Comparison graph of Precision, Recall, and F1-score across all models.

Performance of individual models for compari- son, the performance of all the individual models in terms of validation and test accuracy (left) and macro-averaged F1-score (right) are shown in Figure 4.18. Overall, we observe that ResNet50 consistently

outperformed other models with accuracy and F1-score >0.90 in both validation and test sets, highlighting its robustness and powerful capacity. MobileNetV2 came closely next, preserving accuracy as well as F1-scores at around 0.88, showing an optimal balance between computational cost and prediction accuracy. DenseNet12119 and EfficientNetB3 did similarly well, and both models levelled off at ˜0.84 accuracy and F1-score. Although these results are lower than ResNet50 and MobileNetV2, they indicate fair predictive ability among classes. Our MiniMedNet despite reporting the lowest accuracy (0.77) and F1-score (0.77) demonstrates that clinically acceptable performance can still be obtained from extremely light-weight architectures (~32.5k parameters). Closeness of the validation and test bars in all models influences that the models did not suffer from a large amount of overfitting and no individual architecture was trained with excessive regularization.

Altogether, these comparative results show the trade-off between the high-parameter models (ResNet50, which achieved highest performance) and the light-models (MiniMedNet, which presented worse accuracy but a good efficiency). MobileNetV2 emerges as a trade-off between efficacy and accuracy, with ensembles improving the results beyond individual models.

4.4 Discussion

The authors emphasize their weakness points and show the potential of each model in the experimental tests. Among the pretrained models, ResNet50 always gave the best accuracy of 90.50% and wellbalanced precision, recall, and F1-score. Its deep residual connections facilitated the flow of gradient and robust feature learning ability, which further enjoys its lower representation of relative entropy, and consequently led to its great discriminative power for visually similar classes (e.g., Polyp vs. Polyp Normal). In contrast, DenseNet121 also parameter-efficient (~7M) had a poor classification performance (83.92%) and showed confusion between GERD and GERD N- and GERD Normal. EfficientNetB3 also achieved a fair performance (84.30%), above average results were obtained for recall but it had more variation in training epochs reached than ResNet50. 42 with a

lightweight architecture (2.26M parameters), is a good trade-off between efficiency and accuracy, and has obtained the accuracy of 88.21%, which are very applicable for resource-constrained applications.

The MiniMedNet proposed as a tailored lightweight CNN, having just ~32.5k parameters, has proved that extremely resource-economical architectures are featured in medical imaging. Despite being the least accurate with a value of 77.09%, its extremely low number of parameters demonstrates its relevance and applicability in low-resource clinical settings, e.g., rural hospitals and portable diagnostic systems. Crucially, the ablation study cements the central intuition that reasonable accuracy can be obtained even with extremely small models if good architectural and optimization choices are made.

Overall robustness and generalization performance of the ensemble methods was enhanced. Soft voting ensembles (equal-weighted and validation-weighted) of

the proposed method obtained accuracies of 91.00%, and exhibited similar trends to the one observed with the individual models exploiting their complementary characteristics. The stack ensemble with a logistic regression meta-learner yielded the highest overall score of 92.92% accuracy and close to perfect precision and recall balance for all classes. This indicates the capability of our ensemble learning approach in improving generalizability and eliminating class-specific misclassifications. But ensembles also bring higher computational and memory burdens that make them impractical in real-world operational settings, unless you have access to hardware-supported acceleration.

We used Grad-CAM visualizations to evidence the discriminative regions for predictions of the model as a form of interpretability validation. These heatmaps further demonstrated that the pretrained models and MiniMedNet uniformly attended to areas based on clinical importance such as mucosal patterns, polyp surfaces, and reflux induced lesions. For instance, in Polyp cases, polyp borders were highlighted by Grad-CAM, and in

GERD cases, the model emphasized abnormal esophageal lining. This visual interpretability, in addition to construct clinical trust and demonstrates that the models are learning meaningful representations and not just spurious features!

Taken together, the discussion of results hints at a number of trade-offs. High-performing larger architectures, such as ResNet50, are significantly computationally intensive and hence not well-suited for deployment in resource-limited settings. Some lightweight models such as MobileNetV2 and MiniMedNet are have lower accuracy but offer a light-weight, and fast network. Ensemble methods outperform all set of single models and this comes with an additional computational expense and complexity. Finally, the introduction of Grad-CAM creates a interpretability layer that reinforces the clinical implications of the framework. Collectively, our results highlight that the choice for a model in medical AI-systems should optimize accuracy, interpretability, and computational feasibility with respect to the use case.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

The incorporation of deep learning in gastrointestinal endoscopy image classification can potentially revolutionize health care and society as a whole. AI categorization systems could offer consistent, fast and accurate diagnostics of health conditions, such as gastroesophageal reflux disease (GERD) and gastrointestinal polyps. These approaches attenuate the diagnostic lag and promote early-stage diagnosis and medical intervention, which can greatly benefit patient survival and life quality. In practice, these tools may act as decision support systems for aiding physicians in making diagnosis easier and reducing human errors. Socially, one of the clear advantages is democratization of diagnostic expertise. In most developing or rural areas, the availability of a trained gastroenterologist along with advanced diagnostic facilities is limited. AI-enabled tools can help fill this void by delivering expert-level decision-making support in under-resourced healthcare domains. This also has the potential to decrease of the disparity of healthcare levels in different areas and different populations.

In addition, deep learning systems can help in the optimization of the diagnostic process and efficiency, potentially leading to a decreased cost of delivering healthcare. Invasive diagnostic techniques are also unnecessary, and by providing accurate screening, precancerous polyps can be detected before they develop into costly late - stage treatments. With time, broad use of AI applications could be the key to building a more robust public health infrastructure, more successful disease screening programs and better outcomes for groups of many thousands of patients.

5.2 Impact on the environment

Notwithstanding the clinical advantages, deep learning also has environmental issues that need to be taken into account. Training big models as ResNet50, EfficientNetB3, DenseNet121 -- demands a lot of computing power, usually higher end GPUs (And high consumed electricity) This has a negative impact on greenhouse gases and results in a larger carbon footprint. Ensemble methods lead to improved prediction performance, but also introduce additional computation and training expense.

Nevertheless, there are tactics for addressing these environmental setbacks. First, Small neural networks like MiniMedNet, with only ~32.5k parameters, indicate that competitive accuracy level could be achieved without vast amount of computational resources. These models drastically reduce the amount of power and hardware, serving as a more environmentally-friendly alternative. Second, some algorithmic advancements

like effective training strategies, transfer learning and model distillation offer choices to alleviate the computational burden.

AI-based diagnostics may indirectly contribute to environmental sustainability by limiting unnecessary duplicate diagnostic procedures or consumables as well as patient transportation to healthcare facilities. For instance, AI driven remote diagnostics could reduce travel emissions without sacrificing healthcare quality. In future, the combination of renewable energy in data centers, as well as efficient deployment on mobile devices will be crucial in assuring the alignment of medical AI with global sustainability objectives.

5.3 Ethical Aspects

The application of deep learning in GI diagnoses even has ethical implications, which must be thoughtfully considered for prudent use. Most notable among these is patient confidentiality and security. Strict anonymization, encryption, and compliance with laws, regulation such as

GDPR, and standards such as HIPAA is required to protect the patient rights. Yet another important

ethical issue is algorithmic transparency and interpretability. Before we can practice with AI systems on patients, we need to know how to interpret the products of AI and trust those systems as well. In the present work, we generated Grad-CAM heatmaps to visualize which parts of the image are particularly relevant in the decisions. This transparency is necessary to improve clinician trust and to prevent the "black-box" dilemma in AI-assisted healthcare. Dataset bias is another serious threat. If certain demographic groups are underrepresented in the training data, the model's predictions will fail to generalize, resulting in unfair healthcare outcomes. Therefore, maintaining equity and inclusivity in dataset creation and validation is an ethical mandate. There also has to be an accountability framework: Who is liable when an AI system provides a wrong diagnosis must be clearly laid out. In the end, ensuring the ethical integrity of new health technologies

must be a multidisciplinary effort involving clinicians, developers, regulators, and patient advocacy organizations. Then, and only then, can these systems be safely deployed in real-world healthcare.

5.4 Sustainability Plan

In the longer term, it is important to make sure that AI systems in medical imaging continue to develop in an adaptable, ethical, and sustainable way. It is therefore required to constantly evaluate and adapt data process algorithms to avoid them becoming obsolete and remain valid in a changing clinical setting. Models need to be retrained or adapted to new imaging modalities, patient cohorts, or disease manifestations. Capacity-building in healthcare is essential. AI outputs can, however, elevate the level of clinical decision making when clinicians are empowered through education to be critical of them. Finally, a multifaceted ecosystem combining hospitals, research organizations, technologists, insurers and patient advocacy groups will be instrumental in maintaining momentum. They could develop joint standards, regulations and access provision

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

In this research, we investigated using deep learning for automatically classifying gastrointestinal endoscopy images into either GERD, GERD Normal, Polyp and Polyp Normal. A selection of transfer models EfficientNetB3, ResNet50, DenseNet121 and MobileNetV2 were trained and validated using a curated dataset of 24036 images. Ensemble strategies (ensemble methods) and a dedicated low-weight architecture, MiniMedNet, were also tested in order to achieve an optimal trade-off towards both performance and computational efficiency. The models are evaluated with accuracy, precision, recall, and F1-score, and interpretability through Grad-CAM visualizations. The results showed that ResNet50 consistently performed the best for single models, and the combined models showed further improvement. Although MiniMedNet had fewer parameters, it suggested that resource-efficient architecture for clinical deployment is possible.

6.2 Conclusions

The findings of this study suggest the superiority of transfer learning and ensemble models for the task of gastrointestinal disease classification. ResNet50 obtained the best single-model performance, proving its discriminative power against visually complex classes including polyps and GERD lesions. The performance of classification can be further improved by using ensemble methods, in particular, stacking, which indicated the importance of model fusion. The results of the introduction of MiniMedNet imply that lightweight models can attain clinically acceptable performance with great reduction in computational expense, which indicates their potential for implementation in resource-limited scenarios such as healthcare. Interpretability of Grad-CAM supported a higher practitioner confidence by suggesting alignment between the model's focus areas

and clinically relevant features. In sum, the work validates the importance of a trade-off between accuracy, efficiency, and interpretability in facilitating the take-off of AI in medical imaging.

6.3 Implication for Further Study

The results of this paper suggest a number of directions for future research. Generalization across multiple clinical settings and individual facilities may be increased by augmenting the dataset with multi-center and multi-device endoscopy images. Future work could also consider hybrid models that combine the temporal video-based approach; since endoscope procedures often generate continuous video instead of the still images. Moreover, additional improvement for lightweight models, e.g., MiniMedNet, by neural architecture search or pruning may achieve even more efficient designs without losing much performance. Other explorations of interpretable frameworks beyond Grad-CAM could similarly contribute to offering clinicians with "cleaner" views on AI predictions, promoting fairness, accountability and transparency. Lastly, clinical translation and validation in close partnership with clinicians is paramount to close the gap between the performance observed in experimentation and actual deployment, to guarantee that AI is robust, ethical and enhances patient care.

REFERENCES

[1] C. P. Gyawali and R. Fass, "Management of gastroesophageal reflux disease," *Gastroenterology*, vol. 154, no. 2, pp. 302–318, 2018.

[2] D. K. Rex, C. A. Johnson, J. C. Anderson, et al., "American College of Gastroenterology Guidelines for Colorectal Cancer Screening 2009," *Am. J. Gastroenterol.*, vol. 104, no. 3, pp. 739–750, 2009.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] J. Urban, T. Tripathi, T. Alkayali, et al., "Deep Learning for Real-Time Detection of Colorectal Polyps in Colonoscopy Videos," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[6] J. Urban et al., "Deep Learning for Real-Time Detection of Colorectal Polyps in Colonoscopy Videos," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018.

[7] P. Wang et al., "Artificial Intelligence Aided Colonoscopy for Adenoma Detection: A Randomized Clinical Trial," *Gut*, vol. 68, no. 10, pp. 1813–1819, 2019.

[8] J. Bernal et al., "Deep Convolutional Neural Networks for Polyp Detection in Colonoscopy," *Medical Image Analysis*, vol. 31, pp. 219–235, 2016.

[9] K. Pogorelov et al., "Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in *Proc. MMSYS*, 2017.

[10] C. Li et al., "Automatic Classification of Gastroesophageal Reflux Disease from Endoscopic Images Using Deep Learning," *J. Gastroenterology*, vol. 55, pp. 1–10, 2020.

[11] Y. Zhang et al., "Hybrid CNN-SVM Model for GERD Image Classification," *Computers in Biology and Medicine*, vol. 127, 2020.

[12] F. Xie et al., "Fine-Tuning of ResNet and EfficientNet for Gastroesophageal Reflux Disease Diagnosis," *IEEE Access*, vol. 9, pp. 1–12, 2021.

[13] A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[15] D. Lee et al., "Comparative Study of Lightweight CNNs for GI Endoscopy Classification," *Computer Methods and Programs in Biomedicine*, vol. 196, 2020.

[16] X. Xu et al., "Ensemble Learning for Gastrointestinal Image Classification," *Journal of Biomedical Informatics*, vol. 110, 2020.

[17] Y. Zhang et al., “Stacking Ensemble of CNNs for Radiology Image Classification,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, 2020.

[18] R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *Proc. ICCV*, 2017, pp. 618–626.

[19] Y. Shin et al., “Application of Grad-CAM to Colonoscopy Images for Polyp Detection,” *Endoscopy International Open*, vol. 7, no. 12, 2019.

[20] M. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” in *Proc. KDD*, 2016, pp. 1135–1144.

[21] M. Misawa et al., “Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience,” *Gastroenterology*, vol. 154, no. 8, pp. 2027–2029, 2018.

[22] D. Byrne et al., “Real-time Differentiation of Adenomatous and Hyperplastic Diminutive Colorectal Polyps Using Deep Learning,” *Gut*, vol. 68, no. 12, pp. 2245–2253, 2019.

[23] D. Byrne et al., “Clinical Validation of an AI-Based Polyp Detection System in Colonoscopy,” *The Lancet Gastroenterology & Hepatology*, vol. 4, no. 3, pp. 181–187, 2019.

[24] D. Byrne et al., “Interpretable AI-Assisted Colonoscopy Reduces Miss Rates of Diminutive Adenomas,” *Endoscopy*, vol. 52, no. 10, pp. 857–865, 2020.

[25] Y. Mori et al., “Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study,” *Annals of Internal Medicine*, vol. 169, no. 6, pp. 357–366, 2018.

[26] M. Yamada et al., “Development of a Real-Time Endoscopic Image Diagnosis Support System Using Deep Learning Technology,” *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357–363, 2019.

[27] S. Ali et al., “Shallow Convolutional Neural Networks for Real-Time Polyp Detection,” *Healthcare Technology Letters*, vol. 6, no. 5, 2019.

[28] S. Thambawita et al., “GAN-based Data Augmentation for Improving Gastrointestinal Disease Classification,” *IEEE Access*, vol. 8, 2020.

[29] M. Hossain et al., “Gastrointestinal Endoscopy Image Dataset for GERD and Polyp Classification,” *Mendeley Data*, v3, 2024. DOI: 10.17632/ffyn828yf4.3.

[30] C. P. Gyawali and R. Fass, “Management of gastroesophageal reflux disease,” *Gastroenterology*, vol. 154, no. 2, pp. 302–318, 2018.

[31] N. J. Shaheen and J. E. Richter, “Barrett’s oesophagus,” *The Lancet*, vol. 373, no. 9666, pp. 850–861, 2009.

[32] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 60, 2019.

[33] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” arXiv preprint arXiv:1712.04621, 2017.

[34] F. Zerbib et al., “Modern diagnosis and management of GERD,” *Nature Reviews Gastroenterology & Hepatology*, vol. 17, no. 8, pp. 493–507, 2020.
[35] N. Vakil et al., “The Montreal definition and classification of gastroesophageal reflux disease: A global evidence-based consensus,” *Am. J. Gastroenterol.*, vol. 101, no. 8, pp. 1900–1920, 2006.
[36] L. Lundell et al., “Endoscopic assessment of oesophagitis: Standardization of reporting and grading,” *Gut*, vol. 45, no. 2, pp. 172–180, 1999.
[37] D. K. Rex et al., “Colorectal polyp classification and management,” *Gastroenterology*, vol. 153, no. 3, pp. 722–743, 2017.
[38] C. Hassan et al., “Post-polypectomy surveillance: ESGE Guideline,” *Endoscopy*, vol. 52, no. 8, pp. 687–700, 2020.
[39] S. J. Winawer et al., “Prevention of colorectal cancer by colonoscopic polypectomy,” *New England Journal of Medicine*, vol. 329, no. 27, pp. 1977–1981, 1993.
[40] M. F. Byrne et al., “Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during colonoscopy using deep learning: A prospective trial,” *Gut*, vol. 68, no. 12, pp. 2245–2253, 2019.
[41] M. Misawa et al., “Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience,” *Gastroenterology*, vol. 154, no. 8, pp. 2027–2029, 2018.
[42] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv:1409.1556, 2014.
[43] G. Huang et al., “Densely Connected Convolutional Networks,” Proc. CVPR, 2017.
[44] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training,” Proc. ICML, 2015.
[45] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization (AdamW),” Proc. ICLR, 2019.
[46] T. DeVries and G. W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout,” arXiv:1708.04552, 2017.
[47] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[48] A. Esteva et al., “A Guide to Deep Learning in Healthcare,” Nature Medicine, vol. 25, pp. 24–29, 2019.
[49] S. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions (SHAP),” Proc. NeurIPS, 2017.
[50] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks (Integrated Gradients),” Proc. ICML, 2017.
[51] J. He et al., “The Practical Implementation of AI in Gastrointestinal Endoscopy,” World Journal of Gastroenterology, vol. 27, 2021.

[52] K. Ding et al., “Gastrointestinal Lesion Detection Using Capsule Endoscopy and Deep Learning,” IEEE Access, vol. 8, pp. 10462–10470, 2020.
[53] H. Nakagawa et al., “AI-assisted Gastric Cancer Diagnosis with Endoscopy,” Endoscopy International Open, vol. 8, no. 1, 2020.
[54] S. Min et al., “Deep Learning in Bioinformatics,” Briefings in Bioinformatics, vol. 18, no. 5, pp. 851–869, 2017.
[55] S. Anwar et al., “Structured Pruning of Deep Convolutional Neural Networks,” ACM JETC, vol. 13, no. 3, 2017.
[56] G. Hinton et al., “Distilling the Knowledge in a Neural Network,” arXiv:1503.02531, 2015.

242-25-009

ORIGINALITY REPORT

22%	19%	13%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Superior Science Higher Secondary School Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Daffodil International University Student Paper	2%
4	arxiv.org Internet Source	1%
5	"Selected Proceedings from the 2nd International Conference on Intelligent Manufacturing and Robotics, ICIMR 2024, 22-23 August, Suzhou, China", Springer Science and Business Media LLC, 2025 Publication	1%
6	web.archive.org Internet Source	1%
7	export.arxiv.org Internet Source	<1%
8	deepai.org Internet Source	<1%
9	doaj.org Internet Source	<1%
10	s-space.snu.ac.kr Internet Source	<1%