Universiteit Leiden

Master Computer Science

[An Evaluation of Data Analysis Techniques in Digital Health Applications]

Name:	[Jing Meng]
Student ID:	[s4045874]
Date:	[22/07/2025]
Specialisation:	[Data Science: Computer Science]
1st supervisor:	[Marco Spruit]
2nd supervisor:	[Bram van Dijk]

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

The increasing usage of smart devices and the development of artificial intelligence have also sped up the growth of the digital health industry. Various digital health applications have created new opportunities for the healthcare industry while generating massive data. Due to the diversity and complexity of medical data, it is necessary to select appropriate data analysis techniques for different applications. This study evaluated the usage patterns of predictive data analysis techniques in digital health applications through a literature review, helping developers understand commonly used data analysis techniques in different medical scenarios and select appropriate methods. Additionally, given the rapid development of deep learning models and their increasing usage in the medical imaging field, we need systematic performance benchmarks to assess whether these advanced techniques can bring real improvements. Therefore, we tested various CNN-based models for glaucoma classification using fundus images on a real-world medical database and set corresponding benchmarks. Those experiments provided evidence for model selection and indicated the trade-offs between model complexity and accuracy in practical applications.

We conducted literature screening with the help of ASReview and systematically analyzed 249 articles from four major databases. We categorized the articles by data formats used in the studies (audio, image, video, and structured numerical data) and analyzed the distribution of predictive data analysis techniques. We found that traditional machine learning methods are more suitable for structured numerical data, audio data analysis relies on feature engineering, image data analysis largely depends on CNN architectures, and video data analysis often requires computer vision tools. Additionally, we conducted comparative experiments on the EyePACS-AIROGS-light-V2 glaucoma dataset, evaluating four groups of models: classic stacked convolutional neural networks (CNNs), two-stage transfer learning approaches, end-to-end transfer learning approaches, and hybrid methods combining deep learning with traditional machine learning classifiers. Models using end-to-end transfer learning strategies achieved the best performance. The ResNet50 based model reached 91.95% accuracy, 94.55% recall, and 0.9704 AUC-ROC. Lightweight models (MobileNetV2 and MobileNetV3-Small) also performed well when using end-to-end training and achieved accuracy higher than 91%.

Our study shows the importance of choosing data analysis techniques according to the data format and medical scenarios, providing a reference for application developers in selecting suitable analysis techniques. In addition, through multiple comparative experiments on image data, we also set baselines and proved the potential of lightweight models such as MobileNetV2 and MobileNetV3-Small, providing support for deploying predictive analysis tools with constrained resources in the trend of mobile healthcare.

Keywords: digital health; predictive analysis; artificial intelligence; machine learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The development of the healthcare industry is closely related to people's quality of life. In recent years, the widespread of smart devices such as smartphones and smartwatches and the rapid development of artificial intelligence have changed the way people access healthcare services (Bajwa et al., 2021). Digital health applications provide new opportunities for the development of the healthcare industry (Al Kuwaiti et al., 2023).

Digital health applications such as telemedicine and mHealth can make it easier for residents in remote areas to get healthcare (Wosik et al., 2020; Peyroteo et al., 2021). The development of wearable devices and IoT technology can improve patient's quality of life with chronic diseases while reducing the burden on healthcare providers and increasing the efficiency of healthcare institutions (Tan et al., 2024). In addition, the COVID-19 pandemic has also made people realize the advantages of telemedicine and accelerated the development of digital health, which has also generated a large amount of medical data in the process (Wosik et al., 2020; Tan et al., 2024). The need to analyze medical data also drives the development of data-driven digital health applications (Al Kuwaiti et al., 2023).

However, due to the complexity of medical data, there may be limitations, such as the data being in different formats, data missing (Sedlakova et al., 2023), and strict privacy requirements (Price and Cohen, 2019). Thus, data analysis in the medical industry faces huge challenges. In addition, since the medical industry has a direct impact on the safety of users, any digital health service needs to be evaluated before implementation to ensure that it will not cause harm to users. Therefore, unlike technological innovation in other industries, the speed of technique innovation in the digital health industry is usually faster than the speed of formulating corresponding evaluation guidelines (Mathews et al., 2019). This gap between innovation and validation has led to uncertainty about the effectiveness, safety, and reliability of many digital health applications, and many techniques have not yet been clinically validated (Guo et al., 2020). Only through systematic evaluation, taking into account the effectiveness, stability, and operability of various digital analysis techniques in actual medical settings, can we identify the most promising development directions in this large and rapidly developing field (Guo et al., 2020).

Digital health applications are diverse and numerous, and the data analysis techniques involved are even more complex. In order to explore this area more efficiently, it is necessary to classify these applications first. In 2023, the World Health Organization (WHO) refined the detailed taxonomy of Digital Interventions, Services, and Applications for Health (DISAH) (Organization, 2023). We will use this frame to categorize digital health applications and focus on the category of "Digital Health Interventions for Data Services", particularly those related to "Automated analysis of data to generate new information or predictions on future events".

Applications in this category focus on the use of artificial intelligence, including machine learning, to conduct predictive analytics. These applications can directly improve patients' quality of life and have lots of relevant research resources and publicly available data for validation. In general, they provide an ideal entry point for studying data analysis techniques within digital health. Therefore, this paper focuses on investigating the models and methods used for predictive analysis in digital health applications.

This study can provide a reference for digital health application developers and medical service providers to help them choose appropriate data analysis techniques, especially the predictive models for different digital health applications, and ultimately truly improve the quality of life of every medical service user.

1.2 Problem statement

The research question we explored in this study is: Which predictive data analysis techniques are most commonly adopted in digital health applications, how are these techniques related to different types of medical data, and what are the distribution patterns of various predictive data analysis techniques in real medical environments?

Currently, a wide range of techniques have been applied in the healthcare industry, from traditional machine learning techniques such as Support Vector Machines (SVM) and Random Forests to deep learning models like Convolutional Neural Networks (CNN). Additionally, the formats of medical data in the real world are diverse, such as images collected by medical devices (e.g., CT scans) to detect joint lesions or numerical laboratory tests results to evaluate a patient's blood glucose. Different data formats may require distinct data preprocessing procedures, feature extractions, and selection methods. Furthermore, many data analysis techniques that perform well in laboratory environments may face challenges in real-world healthcare scenarios, such as computational resource constraints, data quality issues, and user acceptance. This gap between theory and practical application can also lead to situations where cutting-edge techniques may not be suitable for real-world digital health applications. Therefore, a framework is needed to help digital health application developers understand the characteristics of data that may be generated in real-world healthcare scenarios and the appropriate data analysis techniques corresponding to them.

Based on this background, we can identify two specific problems:

1. What is the current usage frequency and the trend of various data analysis techniques in the digital health field?

2. What is the most suitable matching between a specific digital health field (such as diagnosis of a specific type of disease) and the types of data that need to be collected and the predictive data analysis techniques that need to be applied?

1.3 Aims and objectives

This study aims to evaluate the current usage of data analysis techniques in different types of data (structured numerical data, image, audio, and video) in digital health applications through systematic literature analysis, identify the advantages and applicable scenarios of various techniques, and provide suggestions for techniques selection for digital health applications. In addition, we select a real-world medical image dataset to test the applicability of multiple predictive data analysis techniques in this specific data format, thereby providing a partial validation of our proposed framework in the context of image data.

1.4 Summary of contributions and achievements

Our study has two main contributions. Firstly, we conducted a systematic literature review using ASReview, analyzing the usage patterns of various predictive data analysis techniques according to the data formats used in the research. Secondly, in order to set systematic performance benchmarks for CNN-based deep learning techniques in medical image analysis and to test whether complex CNN variants and different learning strategies can provide performance improvements in practical digital health applications, we selected several representative CNN variants to conduct comparative experiments on the EyePACS-AIROGS-light-V2 dataset and set benchmarks, providing references for designing image-based diagnostic systems.

1.5 Organization of the thesis report

Our thesis has six chapters: Introduction, Literature Review, Methodology, Implementation and Results, Discussion and Analysis, Conclusions and Future Work.

Chapter 1 introduces the background and research questions of this study. Chapter 2 presents a comprehensive literature review of predictive data analysis techniques in digital health applications. We categorize the research according to the data formats used (audio, image, video, and structured numerical data) and analyze the usage patterns of various predictive data analysis techniques across different medical domains. Chapter 3 describes the dataset (EyePACS-AIROGS-light-V2), the four groups of models used for comparative experiments, and the evaluation metrics used in this study. Chapter 4 presents the implementation details and experimental results for all four experimental groups, including comprehensive performance metrics (accuracy, precision, recall, F1-score, and Area Under the Curve (AUC)). Chapter 5 discusses the experimental results and the limitations of the current experiments. Chapter 6 summarizes the main findings and possible directions for future research.

Chapter 2

Literature Review

2.1 Review Literature with ASReview

In order to get a comprehensive view of this research area, we selected four databases in computer science and medicine: Web of Science, IEEE Xplore, PubMed, ACM Digital Library, and ACM Digital Library, which are the most authoritative databases in the field.
Our search query can be divided into two parts. The first part focuses on data analysis techniques, and the second part limits the search results to the digital health field:
("data analysis" OR "machine learning" OR "artificial intelligence" OR "deep learning" OR "predictive analytics" OR "statistical analysis" OR "data mining")
AND
("digital health" OR "mHealth" OR "eHealth" OR "telemedicine")

Table 2.1: Database Search Results by Platform

Database	**Search Area**	**Document Type**	**Number of Results**
Web of Science	Topic	Article/ Book Chapters	3948
IEEE	Metadata	Conferences/ Journals/ Magazines/ Books	3284
PubMed	Title/ Abstract	Book and Documents/ Classical Article/ Clinical Conference/ Clinical Trial/ Newspaper Article/ Clinical Study	277
ACM	Title/ Abstract / Keyword	Research Article	93
Total			7602
After cleaning and deduplication			7323

As shown in Table 2.1, there were too many articles in the database search phase. We used the open-source active learning tool ASReview to simplify the screening process. ASReview uses machine learning techniques to select and recommend the most potentially relevant papers, significantly reducing the time needed for literature review while maintaining high sensitivity for identifying relevant studies(van de Schoot et al., 2021). This approach enables the efficient management of a large number of potential articles while ensuring the system comprehensively

covers relevant literature.
We stopped screening when we got 10 irrelevant articles in a row. After using the ASReview tool for preliminary screening of the search results, we got 860 articles most relevant to our topic, "Automated analysis of data to generate new information or predictions on future events." In the second round of screening, we did not use the stopping criteria and reviewed all 860 articles to select studies that had already been tested in the real world, as these studies have a greater impact on the real medical world. The final database contained 249 articles.

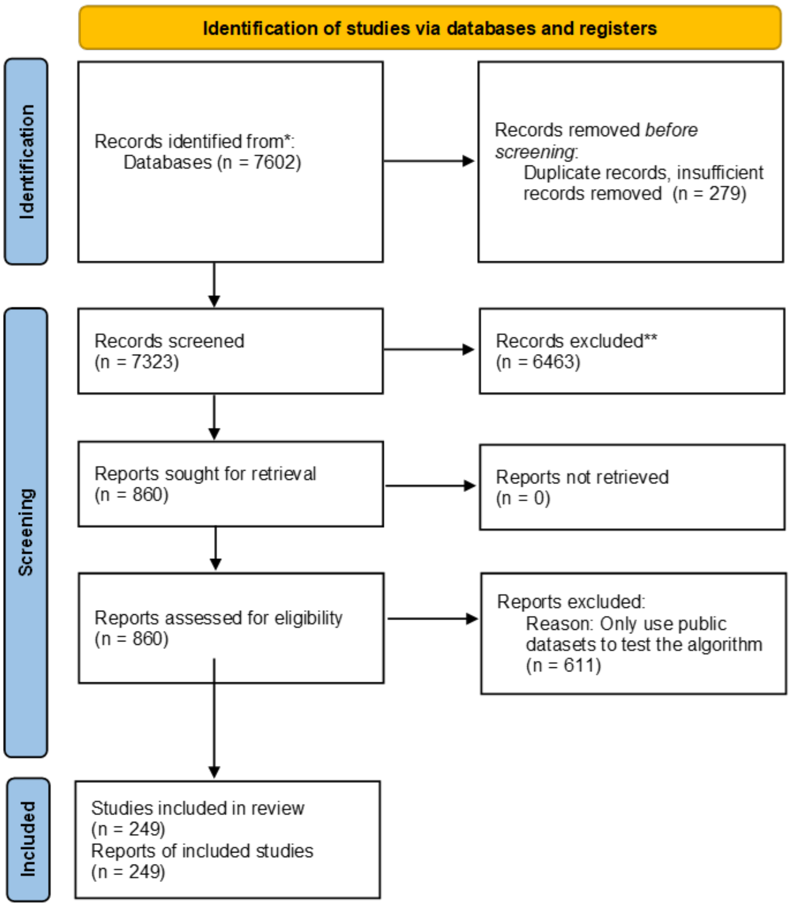

Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n = 7602)
Records removed *before screening*:
Duplicate records, insufficient records removed (n = 279)
Screening
Records screened
(n = 7323)
Records excluded**
(n = 6463)
Reports sought for retrieval
(n = 860)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 860)
Reports excluded:
Reason: Only use public datasets to test the algorithm
(n = 611)
Included
Studies included in review
(n = 249)
Reports of included studies
(n = 249)

Figure 2.1: Literature Screening - PRISMA 2020 flow diagram (Page et al., 2021)

2.2 Main findings and evaluation

We can categorize the predictive data analysis techniques involved in the final literature collection according to the model's structure. Some models may show advantages when processing certain types of data. Figure 2.2 shows the most frequently used models in the relevant articles and the most frequently used models in each data type. There are four main types of data involved in the articles we collected: audio, image, video, and structured numerical data.

2.2.1 Audio

We selected a total of 39 papers that included audio data. In these studies, researchers recorded audio files using mobile phones, wearable sensors, or professional recording devices. The most

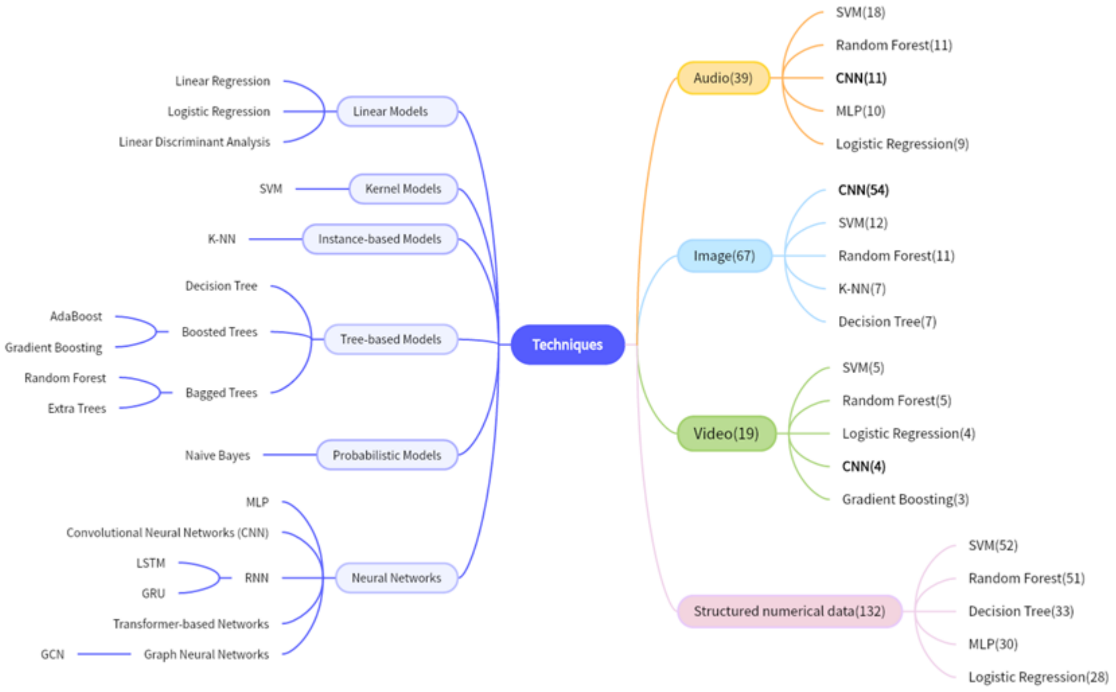


Figure 2.2: Most frequently used models in the relevant articles - Most frequently used models in each data type

common pre-processing method for these audio files is extract features from the time/frequency, cepstral, and wavelet (Radouani et al., 2021*b*) domain. And then we can use feature selection techniques such as ReliefF (Radouani et al., 2021*b*) or openSMILE (Mayr et al., 2025) to identify the main features and then input these features into the machine learning model for analysis to determine the health status of the subjects.

In terms of machine learning techniques, traditional models such as SVM and random forest are widely used, assisting diagnosis of various diseases. For example, detecting Parkinson's disease through voice analysis (Radouani et al., 2021*a*; Zhang et al., 2020; Motin et al., 2022) or monitoring heart condition by analyzing heart sounds (Güven et al., 2021; Narváez et al., 2017). Meanwhile, with the development of the CNN model, some researchers have also applied it to process audio data. We can convert audio signals into spectrograms or Mel-spectrograms, use CNN to extract features from them, and then input these features into machine learning models for analysis (Vatanparvar et al., 2021). Or we can directly use CNN models to classify by learning the features in the spectrograms (Chia et al., 2024; Castillo-Escario et al., 2022). In addition, similar to how we pre-train CNN models on large-scale image datasets and then fine-tune them on small-scale medical image datasets to improve model performance, we can also pre-train CNN models to improve accuracy in audio data analysis (Chia et al., 2024; Hu et al., 2021).

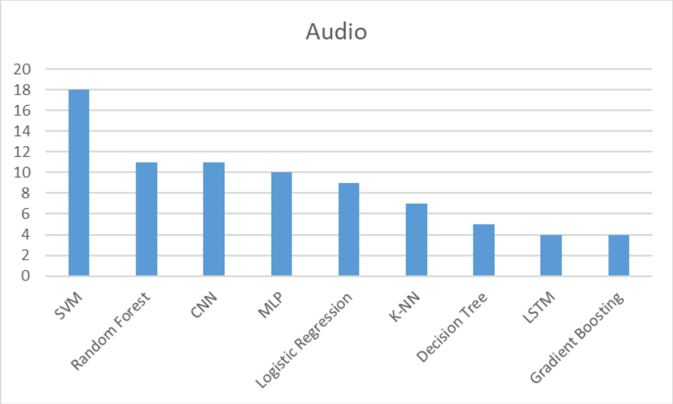


Figure 2.3: Most frequently applied data analysis techniques (Audio)

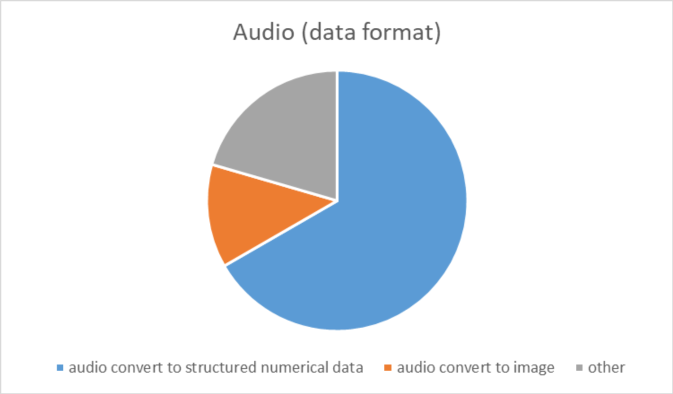


Figure 2.4: Data formats related to audio

Moreover, since audio files can be regarded as changes in sound signals over time, they also have temporal features. Therefore, in theory, recurrent neural networks, especially LSTM (Bi-LSTM) and GRU (Bi-GRU), should also have advantages in processing audio data. The work by Choi et al. uses a BiGRU component that is specifically designed to capture temporal relationships in respiratory sounds to improve their model's performance (Choi et al., 2022). However, RNN-based approaches haven't been widely applied in medical audio data analysis. It will be a direction for future exploration, particularly in developing hybrid architectures that combine CNN feature extraction with the temporal modeling strengths of RNNs.

The audio data in those studies mainly comes from three aspects. First, in the research related to Parkinson's Disease, researchers collect recordings of participants reading specific content to analyze whether they have Parkinson's Disease (Radouani et al., 2021*a*; Zhang et al., 2020; Motin et al., 2022). Secondly, audio can be collected from wearable sensors and stethoscopes when analyzing the respiratory health status of participants (Choi et al., 2022; Chamberlain et al., 2016). In addition, heart sounds are collected during the diagnosis process of cardiovascular diseases, which are mainly used to detect heart rhythm based on heart sounds and then classify health problems such as arrhythmia. For example, the low-cost electronic stethoscope for heart disease detection developed by Jahin et al. uses heart sound recordings and achieves an accuracy rate of 92.48% using artificial neural networks (Jahin et al., 2022).

In general, current audio data analysis in the health field still needs to rely on feature engineering and feature selection and optimization to achieve better diagnosis and predictive analysis results. In the future, we can try to conduct more experiments on neural network methods involving temporal feature processing to explore the application of more complex neural networks in audio data processing.

2.2.2 Image

We selected 67 papers related to image data. Among them, 35 studies used collected images for data analysis directly, while 10 studies transformed the collected image data into structured numerical data and then processed it using machine learning models. Models based on CNN architectures were applied to all 40 publications that used image data directly for diagnosis or prediction(image, audio convert to image, video convert to image). There are various CNN variants in those studies, and lightweight models such as MobileNetV2 (Ngeh et al., 2020) and DenseNet121 (Warin et al., 2021) are popular in applications that have limited resources and require faster calculations. For example, Ngeh et al. focused on developing a low-cost solution for skin cancer detection in rural communities using edge computing. Their approach shows lightweight CNN architectures can maintain diagnostic performance while operating within the computational constraints of edge devices.
Furthermore, many researchers chose models pre-trained on large-scale datasets such as ImageNet to reduce training costs and reduce the impact of limited dataset sizes in real-world experiments (Hwang et al., 2019; Girmaw and Taye, 2025).
In studies that convert image data into structured numerical data, traditional machine learning models such as SVM, Random Forest, and K-NN were widely used. This observation is consistent with our findings in audio data processing. However, these studies often required carefully designed feature engineering. For example, Zhang et al. applied a two-level stationary wavelet entropy (SWE) technique to extract meaningful features from brain images in their study on multiple sclerosis detection using MRI scans. The low-dimensional feature space created by SWE was particularly suitable for distance-based classification methods, enabling the K-NN classifier to achieve an impressive accuracy of 97.94
Furthermore, some researchers also converted data from other formats into image formats for analysis, such as converting audio signals into image representations, which we discussed previously. And video data can be converted into image (sequences), which we will explore in the next subsection.

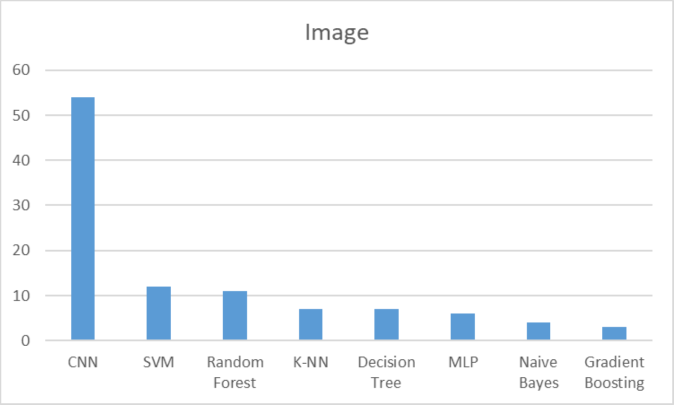


Figure 2.5: Most frequently applied data analysis techniques (Image)

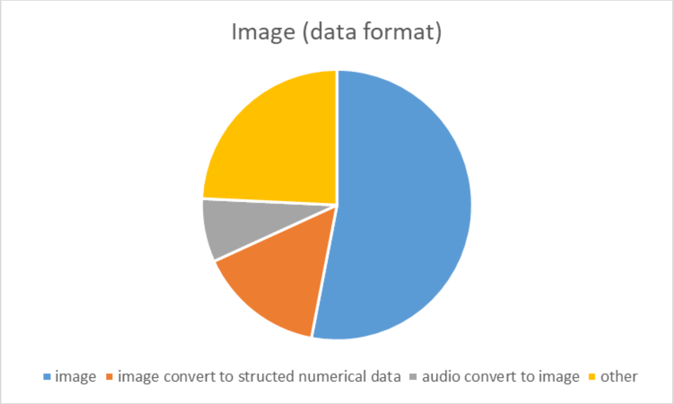


Figure 2.6: Data formats related to image

The image data mainly collected from dermatological and ophthalmological disease diagnosis. Standard dermatological examinations need to use a professional device such as a dermatoscope, while Ophthalmological examinations often rely on fundus images to determine specific disease types. However, with the widespread use of smart devices, it has become possible to collect images using smartphones and other devices.
Overall, CNN is the most important model for analyzing image data in digital health. As smartphones become more popular, using smartphones for preliminary screening real-time diagnostics can significantly lower the costs to accessing healthcare services.

2.2.3 Video

We selected 19 articles that included video data. Among them, 13 studies first converted the video data into structured numerical data and then applied traditional machine learning techniques such as SVM and Random Forest for diagnosis and predictive analysis. This process typically involves using tools like OpenPose and Google MediaPipe for pose estimation and obtaining the coordinates of key points such as joints (Mejía et al., 2022; Guarín et al., 2024; Yang et al., 2021). In addition, it is also possible to directly extract spatial and temporal features using 3D CNNs (Zheng et al., 2022) or convert video data into sequences of images and then apply 2D CNNs (Pourazad et al., 2020) in combination with LSTM for analysis.

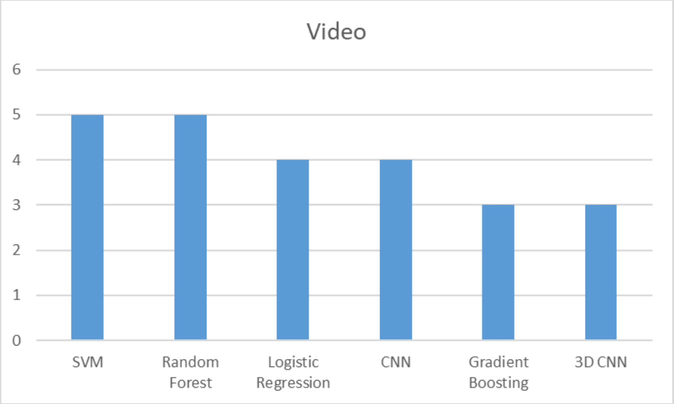


Figure 2.7: Most frequently applied data analysis techniques (Video)

These video data are often collected in studies focused on movement disorders related to Parkinson's disease. The number of studies based on video data is relatively small, indicating that data analysis techniques for digital health applications based on video data are still in the early stages. This is mainly due to the complexity of processing video data, which requires specific computer vision tools and consumes significant computational resources.
In general, analyzing video data in the medical and healthcare field still faces many challenges. However, with the widespread of smartphones, video collection has become much more convenient. In the future, if standardized frameworks for video collection and pre-processing can be built, and with the development of lightweight models, there is great potential for solutions in remote diagnosis and at-home treatment of chronic diseases.

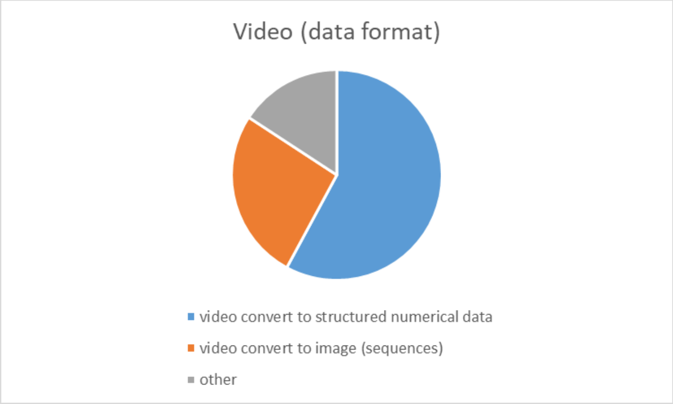


Figure 2.8: Data formats related to video

2.2.4 Structured Numerical Data

We selected 132 articles that used structured numerical data (excluding those that were transformed from images, audio, video, or text). For structured numerical data, traditional machine learning models such as SVM, Random Forest, and Decision Tree are generally applied rather than deep learning methods. This may be due to their lower computational requirements, ease of implementation, and the data's inherent structure, making it easier to process.

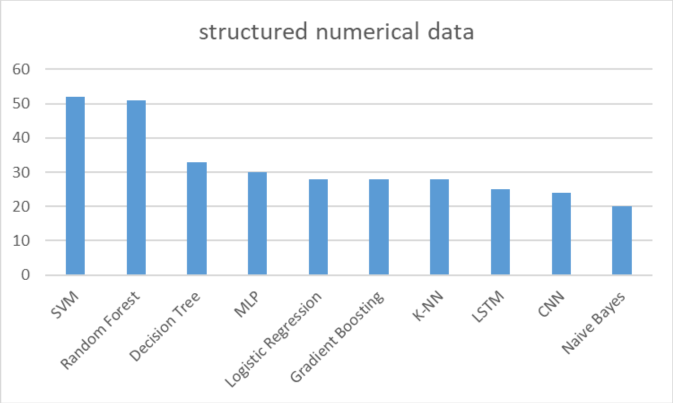


Figure 2.9: Most frequently applied data analysis techniques (Structured numerical data)

Regarding data sources, researchers collected data using mobile devices such as wearable sensors or phone sensors in more than half of the studies. This reflects the trend of mobile medical devices and the integration of healthcare services into users' daily lives. Xiang et al. use mobile phone data for blood pressure prediction and show that daily routine patterns can effectively predict blood pressure (Xiang et al., 2022). Park et al. use sensor-inherited insoles and machine learning to classify abnormal gaits (Park et al., 2024). Fazeli et al. use smartwatch sensor and deep learning to monitor stress levels, showing that physiological signals such as heart rate can predict anxiety in daily life (Fazeli et al., 2022).

Most of these structured numerical data were collected in studies related to human activity recognition, gait analysis and fall detection (Park et al., 2023; Pan and Nan, 2024; Liu et al., 2023; Zhang et al., 2024). The high proportion of such studies in our literature list may also be due to our selection criteria, which required that studies involve real-world experiments rather than only tested on public datasets. The wearable sensors in these human activity recognition studies have, to some extent, lowered the barrier to conducting experiments, making it more possible for such research to be tested in real-world settings. Additionally, data collection in these experiments typically involves only a few dozen participants, each generating a large number of data points by performing multiple types of activities with sensors—another reason why these studies are easier to conduct.

Another source of structured numerical data comes from physiological data obtained in medical examinations. Electrocardiogram (ECG) data is the most common type (Randazzo et al., 2024). As a result, there are a group of studies focusing on cardiovascular health. Kashani et al. developed a lightweight and comprehensive AF detection system. Using only five HRV features with a depth-5 decision tree achieved good performance (98.63% accuracy on AFDB with 99.22% sensitivity) (Kashani et al., 2022). Their work shows signal preprocessing, feature engineering, and model selection can lead to good performance with relatively simple algorithms.

2.2.5 Text

Besides audio, images, videos, and structured numerical data, text is also a common data format in our daily lives. However, among the literature we collected, only five articles used unstructured text data. Only two of them rely entirely on text data. Tariq et al. use the posts from the mental illness-related sections on the Reddit platform to predict mental health state (Tariq et al., 2019), and Soguero-Ruiz et al. use unstructured text medical records from hospitals to predict postoperative complications (Soguero-Ruiz et al., 2016). Both of them use bag-of-words models to extract features after pre-processing and cleaning, and then use traditional machine learning models for predictive analysis. We do not have enough articles for further analysis. It might be because the text related to medical diagnosis contains patient privacy, and annotating medical texts is more challenging, making it difficult to obtain large-scale datasets, which limits the development of related research.

2.3 Summary

By analyzing the selected papers according to the types of data used, we found that different data formats tend to be preferred in the study of different health conditions, and each data type is typically associated with specific suitable data processing techniques. Studies using structured numerical data for analysis are the most common, with more mature data analysis techniques and more standardized analytical methods, usually using traditional machine learning models such as SVM and Random Forest. For audio data, although some researchers tend to convert audio to images and use CNN and other deep neural network models, the combination of feature engineering with traditional machine learning models is still widely used. The application of video data faces greater challenges, with relatively few related studies, as it typically requires computer vision tools and needs complex preprocessing steps. The unstructured text data, such as medical records, faces even more challenges, including patient privacy protection and the time costs of annotation by medical experts. In contrast, research based on medical image data for predictive analysis has higher potential and practical value.

First of all, from the perspective of data collection, the development of cameras in various consumer devices, such as mobile phones, makes it more convenient to take images that can be used for medical diagnosis. Moreover, compared with the high requirements for recording audio and video, taking images is more convenient and feasible. Especially for research on skin-related diseases that mainly rely on the appearance of the lesion, although there are more accurate devices such as dermatoscopes, some researchers have used mobile phones to collect image data and applied them for medical diagnosis research, which greatly reduces the threshold and cost of data collection. For example, Pangti et al. developed a mobile application based on DenseNet-161, using smartphones to take photos with varying camera quality and lighting conditions to prove the possibility of AI-assisted skin disease diagnosis in a real clinical environment (Pangti et al., 2021).

Secondly, from the perspective of data analysis techniques associated with image data, the continuous development of deep learning models provides more practical and efficient options in this field. As we mentioned earlier, image data analysis relies heavily on CNN models, and the emergence of CNN variants in recent years has provided more choices. Particularly, the development of lightweight architectures such as MobileNet has made it more practical to deploy those high-performance image analysis models. The application of pre-trained models and transfer learning techniques is particularly suitable for the relatively small scale of medical datasets. Using large-scale datasets such as ImageNet to provide foundational features for models, enabling them to achieve good performance on relatively small medical image datasets. Additionally, from our previous literature review in section 2.2.1 and 2.2.3, we can also find that image data serves as a commonly used type for data format conversion (Figure 2.4, Figure 2.8). Both audio and video data can be transformed into image data and then benefit from the powerful analytical capabilities of various CNN models. Further proving the potential of image based medical data analysis. Therefore, we should focus on data analysis techniques required for image data-based digital health applications, especially the practical performance of various CNN based techniques under different medical circumstances.

Chapter 3

Methodology

As mentioned earlier, Convolutional Neural Network (CNN) models have been applied widely to assist medical diagnoses based on images. However, there are lots of variations based on CNN, and there are different training strategies that can also lead to performance differences. This experiment aims to conduct a comprehensive comparison of multiple machine learning methods based on CNN in the task of glaucoma detection using fundus images, test the effectiveness of these predictive data analysis techniques, and set performance benchmarks. This comparison will also help future researchers identify the most potential architectures and training strategies for predictive analysis based on medical data in image format. More specifically, this study will test whether more computationally expensive model architectures(Resnet 50) and training strategies(two-stage fine tuning) provide sufficient performance improvements compared to simpler alternatives(classic CNN, Mobilenet V2, Mobilenet V3small; end-to-end training).

3.1 Dataset descriptions

Glaucoma is the second leading cause of blindness worldwide, and early diagnosis is critical to preventing irreversible vision loss. In addition to basic clinical examinations such as intraocular pressure and visual field tests, AI-based analysis of fundus images can offer a convenient solution for large-scale population screening for glaucoma. In the screening process, the label RG (Referable Glaucoma) indicates that the fundus image shows suspicious glaucomatous signs and the patient should be referred to a specialist for further examination or treatment; NRG (Non-Referable Glaucoma) means no obvious signs of glaucoma have been detected, and the patient doesn't need to be referred to a specialist.

We chose the EyePACS-AIROGS-light-V2 dataset, which contains 9,540 standardized color fundus images. It is a balanced subset selected by Kiefer et al. from the large-scale Rotterdam EyePACS AIROGS dataset, specifically designed for machine learning research (Kiefer, 2024). The dataset is divided into a training set (RG: 4,000 images; NRG: 4,000 images), a validation set (RG: 385; NRG: 385), and a test set (RG: 385; NRG: 385). All images are standardized using the CROP method, which proved to be the most effective strategy in their ablation study (Steen et al., 2023). They removed the black background in the fundus images before cropping and resizing, preserving the maximum useful information about the retina.

3.2 Experiments design

3.2.1 Compared models

We selected four groups of representative models for this study. Classic stacked convolutional neural network; Lightweight models represented by MobileNetV2 (Sandler et al., 2018) and

MobileNetV3-Smal (Howard et al., 2019); And a deep, high-performance model represented by ResNet50 (He et al., 2016). In terms of training strategies, we compared two main approaches: Two stages transfer learning and End-to-end training (Yosinski et al., 2014). Finally, we introduced a hybrid method that combines traditional machine learning classifiers with deep learning models as feature extractors to explore more possibilities of predictive data analysis techniques in the medical field.

The first model uses a fully customed convolutional neural network architecture, and the results can be used as the baseline. This convolutional neural network contains five convolutional blocks, each of them having a convolutional layer, batch normalization, ReLU activation function, and max pooling layer. Followed by global average pooling and a fully connected layer for binary classification.

The second group of models uses the two-stage transfer learning approach. We use ResNet50, MobileNetV2, and MobileNetV3-Smallas backbone networks, all of which were pre-trained on the ImageNet dataset. In the first stage, all parameters of the pre-trained backbone are frozen, and only the newly added classifier head is trained. With the learning rate set to 0.0005 and training for 15 epochs. The purpose of this stage is to make our newly added classifier adapt to the mapping relationship from pre-trained features to glaucoma detection labels.

In the second stage, we unfreeze the last 30 layers of the pre-trained network and use a lower learning rate of 0.00005 to fine-tune the models on the EyePACS-AIROGS-light-V2 dataset for 10 epochs. The fine-tuning stage allows the model to adapt to the specific characteristics of the dataset, and a lower learning rate is used to avoid losing the useful representations that the models learned during the pre-training process.

The third group of models uses the end-to-end training strategy. We also use pre-trained ResNet50, MobileNetV2, and MobileNetV3-Small as backbone networks. However, unlike the second group, we won't use the freeze-unfreeze approach. Instead, all parameters of the entire network are set to be trainable, allowing for gradient updates across all layers. The learning rate is set to 0.0005, and the training is conducted for 15 epochs in total to ensure the networks have sufficient time for comprehensive parameter optimization.

The fourth group of models uses a hybrid approach. We select the best-performing model from the third group as a feature extractor and train it on the current dataset for 15 epochs to fully learn features and generate feature vectors. These extracted features will be used as inputs of the traditional machine learning classifiers to produce the final binary classification results.

3.2.2 Evaluation

We evaluated the predictive performance of each method using accuracy, precision, recall, F1-score, and AUC-ROC (Goodfellow et al., 2016). Since in medical practice, the cost of misclassifying a positive patient as negative (false negative) is usually higher than the opposite, we will tend to choose models with higher recall when other performances are similar to reduce the risk.

	Predicted	
Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 3.1: Confusion Matrix

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

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

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

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

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is calculated as:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{3.5}$$

TPR (True Positive Rate) is equal to Recall and FPR (False Positive Rate) is defined as:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3.6}$$

Through these four groups of systematic comparative experiments, we aim to assess the effectiveness of various predictive data analysis techniques, provide practical insights for developers of digital health applications, and promote the real-world application of artificial intelligence in the early screening of ophthalmic diseases.

Chapter 4

Implementation and Results

We systematically evaluated various CNN based deep learning methods on the task of glaucoma classification using fundus images. From classical CNN to pre-trained models and hybrid methods combining deep learning with traditional machine learning, we explored how model architecture and training strategies affect the results of medical image analysis tasks.

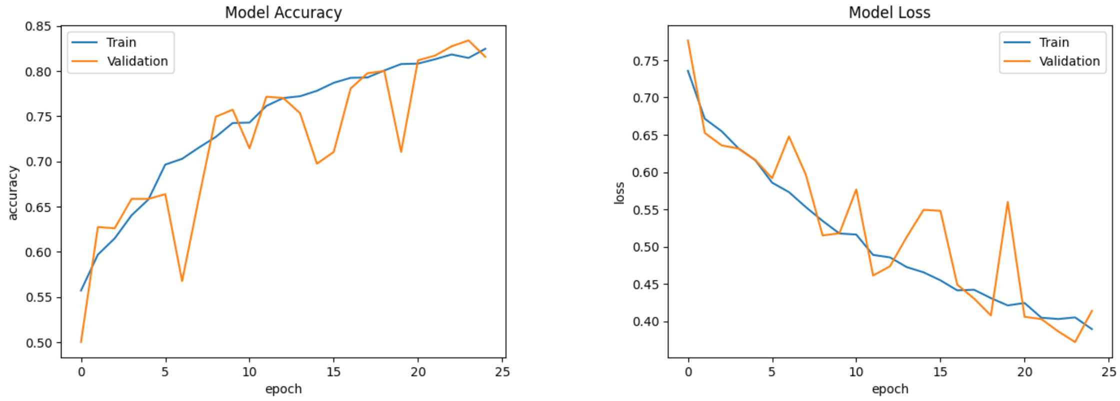
Table 4.1: Performance Comparison of All Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Classic CNN					
Classic CNN	83.12%	79.86%	88.57%	83.99%	0.9143
Two-Stage Transfer Learning					
ResNet50	89.22%	88.72%	89.87%	89.29%	0.9362
MobileNetV2	83.64%	82.95%	84.68%	83.80%	0.9193
MobileNetV3-Small	84.81%	83.50%	86.75%	85.10%	0.9228
End-to-End Transfer Learning					
ResNet50	92.34%	89.37%	**96.10%**	**92.62%**	**0.9756**
MobileNetV2	**92.47%**	**92.69%**	92.21%	92.45%	0.9728
MobileNetV3-Small	90.65%	91.08%	90.13%	90.60%	0.9631
Hybrid Approach					
ResNet50+SVM	87.01%	86.45%	87.79%	87.11%	0.9301
ResNet50+RandomForest	81.43%	78.54%	86.49%	82.32%	0.9084

4.1 Classic stacked convolutional neural network

The classic stacked convolutional neural network uses a progressive feature extraction strategy. After data augmentation (including normalization, horizontal flipping, vertical flipping, and brightness adjustment), the image data will flow into the model. The number of filters gradually increases from 32, 64, 128, 256 to 512. Each convolutional block includes batch normalization and ReLU activation function, followed by global average pooling with dropout regularization. The final output layer uses a sigmoid activation function for binary classification. The model uses the Adam optimizer with an initial learning rate 0.0001 and binary cross-entropy as the loss function. We use an adaptive learning rate scheduling strategy (ReduceLROnPlateau) to prevent overfitting during training. It will reduce the learning rate to

85% of the original value when the loss on the validation set stops improving, with a minimum learning rate limit of 0.00001.

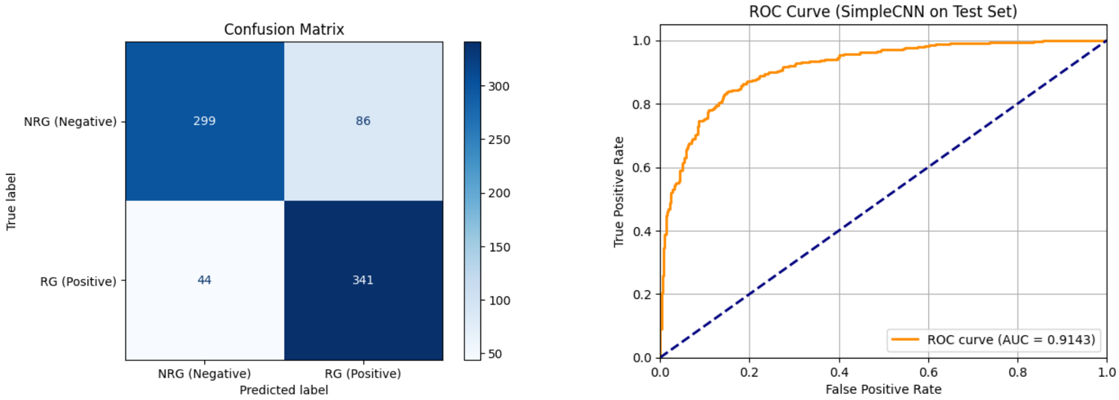


(a) Accuracy (classic stacked CNN)

(b) Loss (classic stacked CNN)

Figure 4.1: Training accuracy and loss of the classic stacked CNN model

As can be seen from figures 4.1a and 4.1b, during the training process, the model's accuracy on the training set improved, and the loss decreased. On the test set, the model achieved an overall accuracy of 83.12%, AUC-ROC of 0.9143, precision of 79.86%, recall of 88.57%, and F1-score of 83.99%. The confusion matrix in figure 4.2a shows that the model tends to predict the image as the positive class (RG), causing 86 false positives and 44 false negatives. The ROC curve 4.2b is obviously convex, which means the model has high sensitivity and specificity under most threshold settings, further proving that classic CNN is already effective on this dataset.



(a) Confusion matrix (classic stacked CNN)

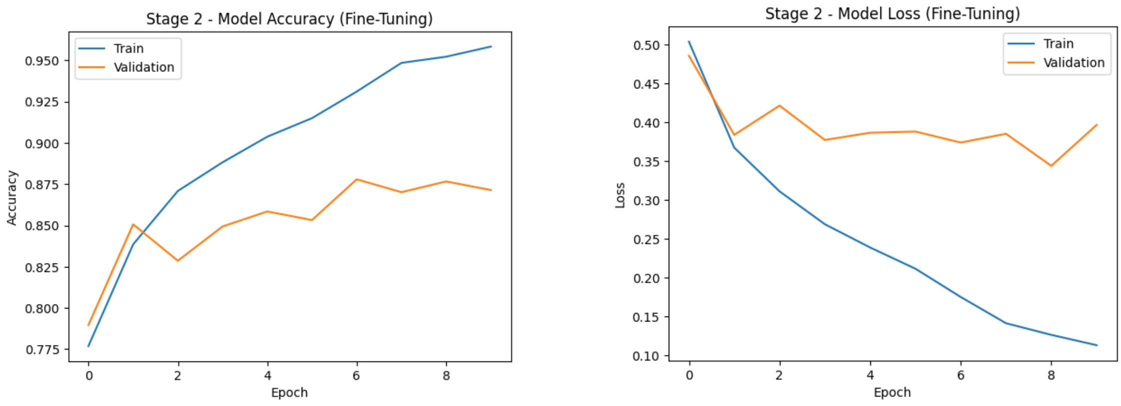
(b) ROC curve (classic stacked CNN)

Figure 4.2: Confusion matrix and ROC curve of the classic stacked CNN model

4.2 Two-stage transfer learning

In the second group of experiments, we used three ImageNet pre-trained models as backbones: ResNet50, MobileNetV2, and MobileNetV3-Small. All three models adopted the same two-stage training strategy to ensure fairness and comparability of the experiments.

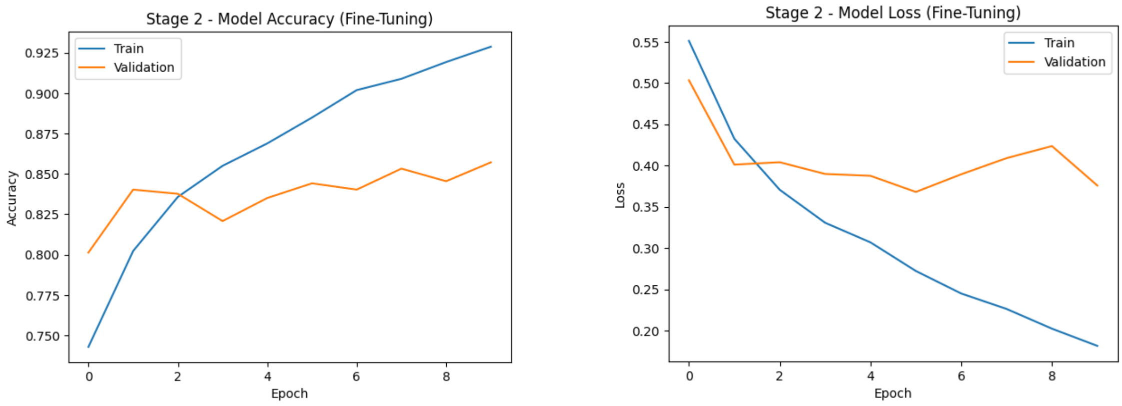
In the first stage, we froze the backbone part of the pre-trained network, removed the original classification head, and added a global average pooling layer and a single sigmoid output node for binary classification. Image data was preprocessed using the preprocess input function specific to each backbone network to ensure input data corresponded with the pre-trained models. We also keep the data augmentation strategies on the training dataset, including horizontal flipping, vertical flipping, and brightness adjustment, to improve model generalization ability. After training for 15 epochs using the Adam optimizer with a learning rate of 0.0005, we proceeded to the second stage, where we unfroze the last 30 layers of the backbone network and reduced the learning rate to 0.00005 for fine-tuning. Both stages employed the same adaptive learning rate decay strategy (ReduceLROnPlateau) as used in the first group of experiments. As we can see in the table 4.1, ResNet50 shows its excellent learning capability on this medical image dataset. MobileNetV2 and MobileNetV3-Small, as representatives of lightweight models, also performed well. Especially MobileNetV3-Small is a choice that balances lightweight with good performance. It has the most stable training process, with the least fluctuation in validation loss, indicating good generalization ability.



(a) Accuracy (ResNet50)

(b) Loss (ResNet50)

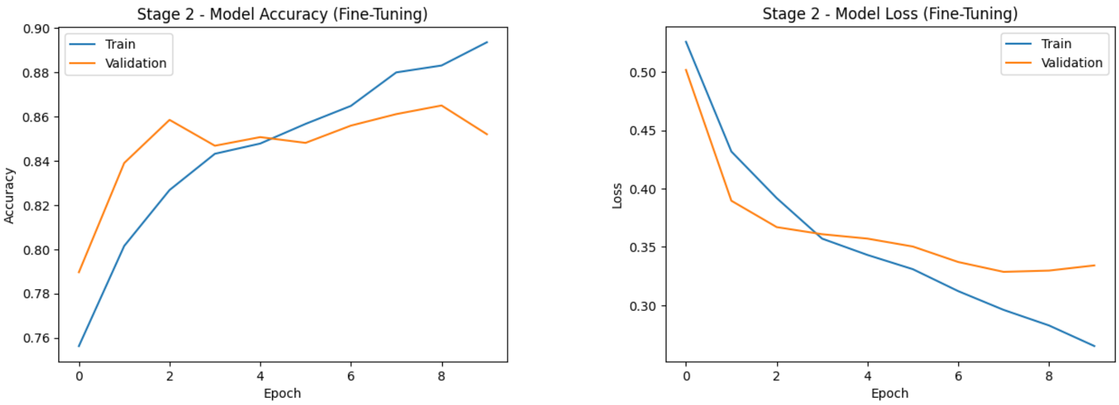
Figure 4.3: Training accuracy and loss (two-stage ResNet50)



(a) Accuracy (MobileNetV2)

(b) Loss (MobileNetV2)

Figure 4.4: Training accuracy and loss (two-stage MobileNetV2)



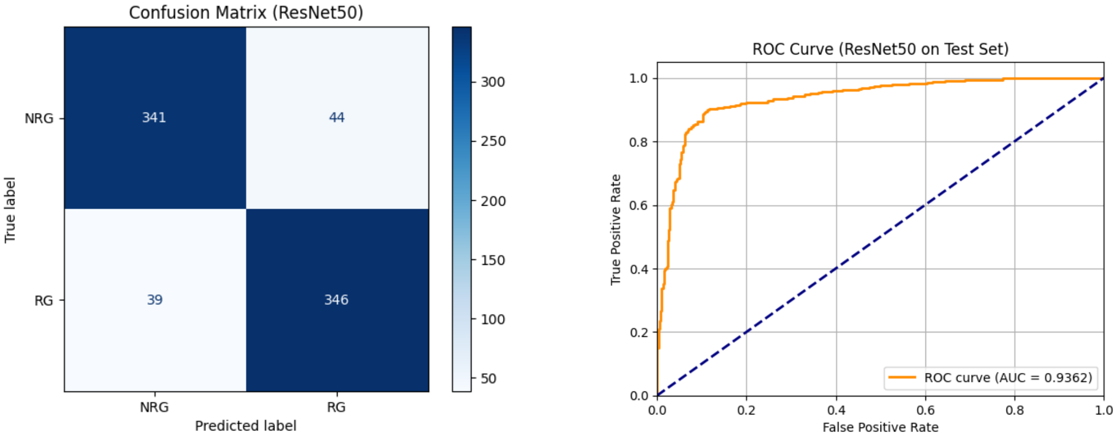
(a) Accuracy (MobileNetV3Small)

(b) Loss (MobileNetV3Small)

Figure 4.5: Training accuracy and loss (two-stage MobileNetV3Small)

On the other hand, as shown in figures 4.3 and 4.4, we should notice that due to their complex structure and the large number of parameters, they have a higher risk of overfitting compared to the other two models. If we want to use them in medical practice, they might require more refined regularization strategies.

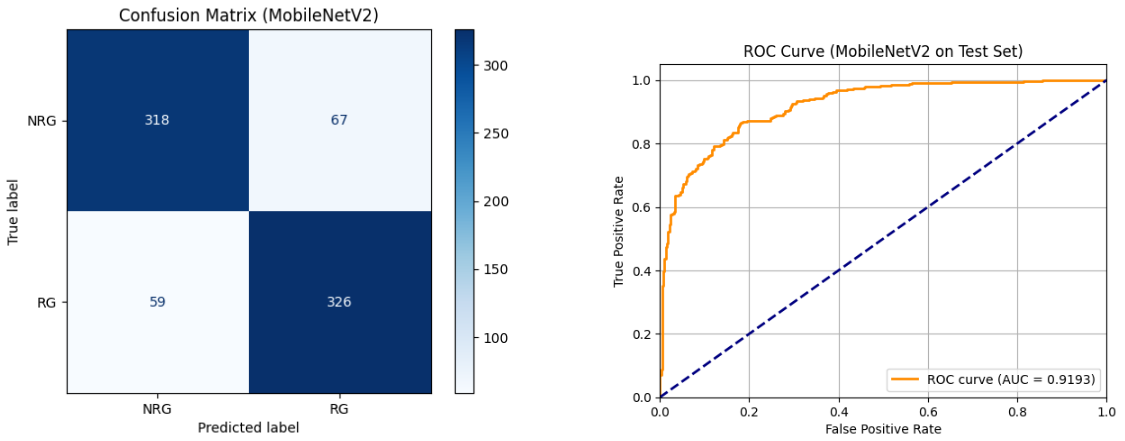
By comparing the confusion matrices and ROC curves of the three models in figures 4.6, 4.7 and 4.8, we can find that ResNet50 has predictive ability in both classes with no obvious bias. It maintains a balance between high sensitivity and specificity across various threshold settings. MobileNetV2, as a lightweight model, has more misclassifications than ResNet50, but its classification is relatively balanced. MobileNetV3-Small is the lightest model, producing 66 false positives and 51 false negatives, indicating a slightly unbalanced performance. However, its AUC value of 0.9228 suggests that the model is still effective.



(a) Confusion matrix (ResNet50)

(b) ROC curve (ResNet50)

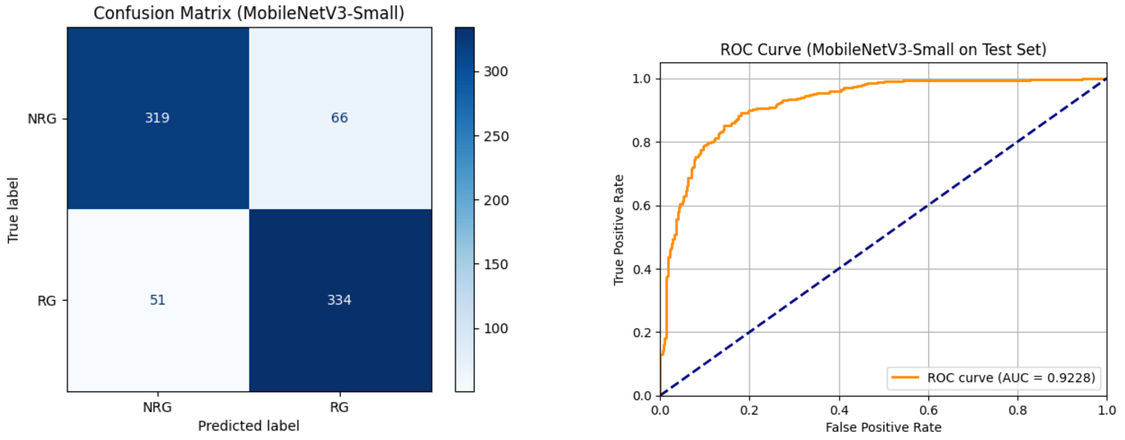
Figure 4.6: Confusion matrix and ROC curve (two-stage ResNet50)



(a) Confusion matrix (MobileNetV2)

(b) ROC curve (MobileNetV2)

Figure 4.7: Confusion matrix and ROC curve (two-stage MobileNetV2)



(a) Confusion matrix (MobileNetV3Small)

(b) ROC curve (two-stage MobileNetV3Small)

Figure 4.8: Confusion matrix and ROC curve (two-stage MobileNetV3Small)

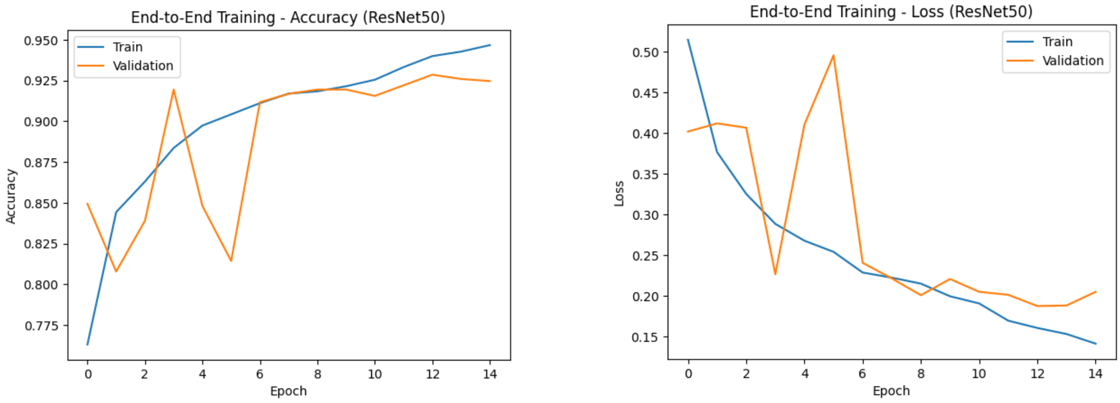
When comparing these three pre-trained and two-stage fine-tuned models with the classical CNN used in the first experiment, the performance improvement is not as significant as expected. This may be because, during two-stage fine-tuning, the lower layers of the models were frozen, preventing them from capturing features that are crucial for glaucoma classification. In the third experiment, we will conduct end-to-end training using the same backbone architectures to test whether they can get those crucial features and better results.

4.3 End-to-end transfer learning

In the third group of experiments, we used the same pre-trained models as in the second group (ResNet50, MobileNetV2, and MobileNetV3-Small) but switched from a two-stage transfer learning approach to an end-to-end training strategy. Therefore, we can evaluate whether directly fine-tuning the entire pre-trained model could achieve better performance in this application scenario. We adopted the same data preprocessing strategy as in the second group, using the preprocess input function specific to each model to ensure consistent input formats, along with the same data augmentation techniques (horizontal flipping, vertical flipping, and brightness adjustment). Unlike the second group, we removed the backbone freezing step and set all layers to be trainable from the beginning. We used a unified learning rate of 0.0005 and the same adaptive learning rate scheduling strategy (ReduceLROnPlateau). Each model was trained for 15 epochs.

As it shown in table 4.1, the accuracy of the ResNet50 model under the end-to-end training strategy is 92.34%, and the AUC-ROC value is 0.9756. From the confusion matrix in the figure 4.12, we can see that ResNet50 misclassified 44 NRG samples as RG and 15 RG samples as NRG. It means that ResNet50 tends to predict the image as the positive class (RG), causing a large number of false positives (44), and the model's performance on the two categories is unbalanced. The accuracy of MobileNetV2 is 92.47%, and the AUC-ROC is 0.9728. The confusion matrix in the figure 4.13 shows that MobileNetV2 misclassified 28 NRG samples as RG and 30 RG samples as NRG. The nearly equal distribution shows that MobileNetV2 has nearly balanced performance on the two categories. Although MobileNetV3Small has the lowest overall accuracy of 90.65%, its AUC-ROC is 0.9631, meaning that the model is still effective. And according to the confusion matrix in the figure 4.14, it misclassified 34 NRG samples as RG and 38 RG samples as NRG. Although the total number of misclassifications (72) is higher, the error distribution is balanced, indicating that it also has nearly balanced performance on the two categories.

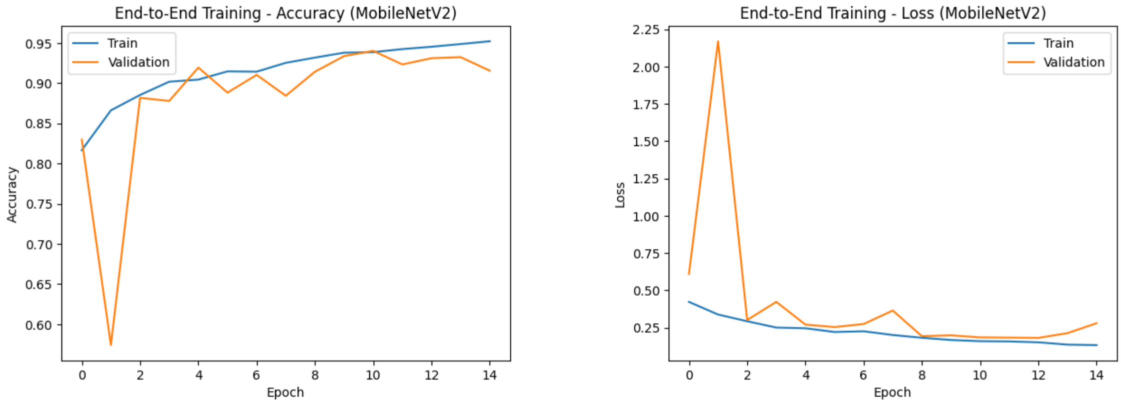
The end-to-end training strategy led to significant performance improvements across all three models. All the pre-trained models now outperform the classical CNN used in Experiment 1, proving the effectiveness of those pre-built CNN variations and their pre-trained weights.



(a) Accuracy (ResNet50)

(b) Loss (ResNet50)

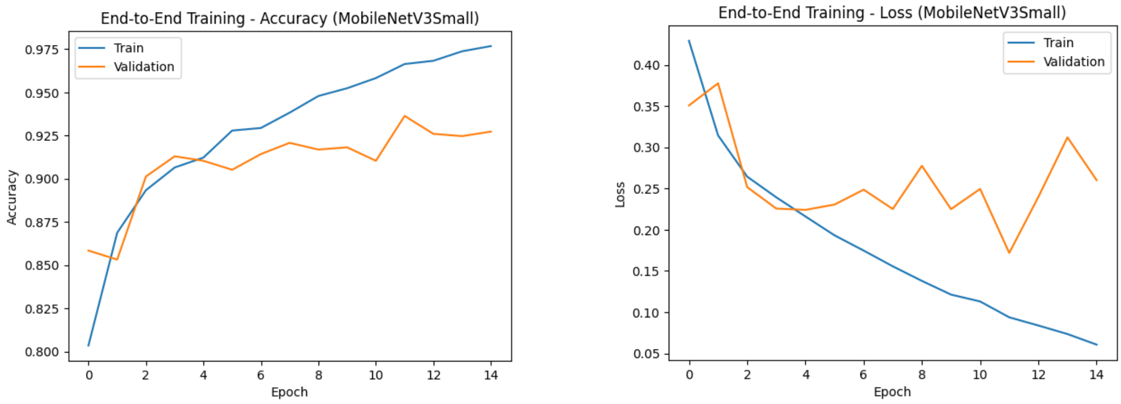
Figure 4.9: Training accuracy and loss (end-to-end ResNet50)



(a) Accuracy (MobileNetV2)

(b) Loss (MobileNetV2)

Figure 4.10: Training accuracy and loss (end-to-end MobileNetV2)

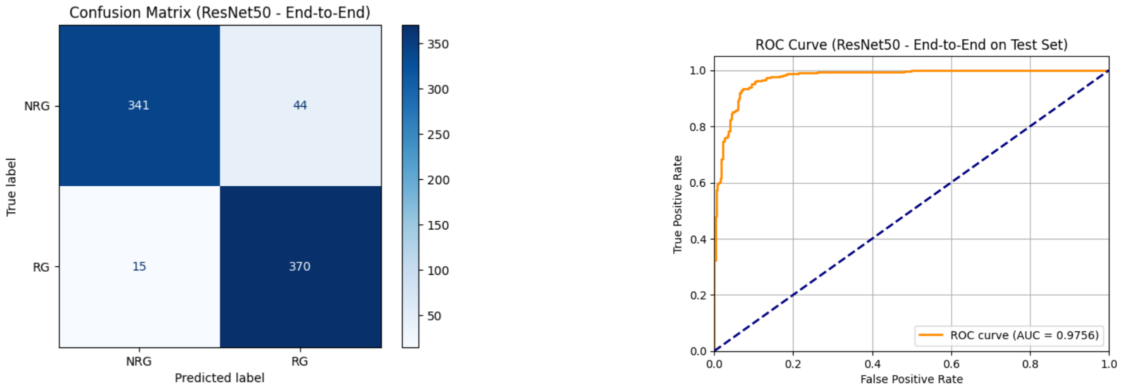


(a) Accuracy (MobileNetV3Small)

(b) Loss (MobileNetV3Small)

Figure 4.11: Training accuracy and loss (end-to-end MobileNetV3Small)

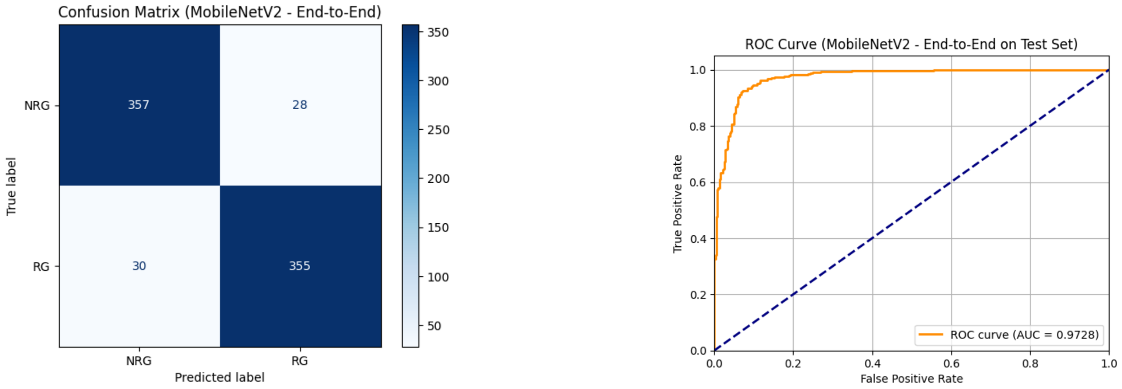
Comparing the results of the second and third groups, we can find that models using end-to-end training strategies actually performed better than those using two-stage fine-tuning on the specific dataset. Considering the dataset we used, the reason might be that glaucoma detection requires attention to features such as optic disc morphology and nerve fiber layer thickness. End-to-end training allows lower-level convolutional blocks to adjust according to the texture and color features of retinal images, while the freezing stage in two-stage learning may limit the model's ability to learn these medical image-specific features, potentially causing feature mismatch in two-stage training. Additionally, due to the small size of the dataset, the original pre-trained weights in the backbone network can provide a good initialization that enables the model to better adapt to the feature distribution of limited data. Two-stage training may disrupt some of the features learned by the backbone network. Therefore, end-to-end training may get better results when dealing with small medical datasets.



(a) Confusion matrix (ResNet50)

(b) ROC curve (ResNet50)

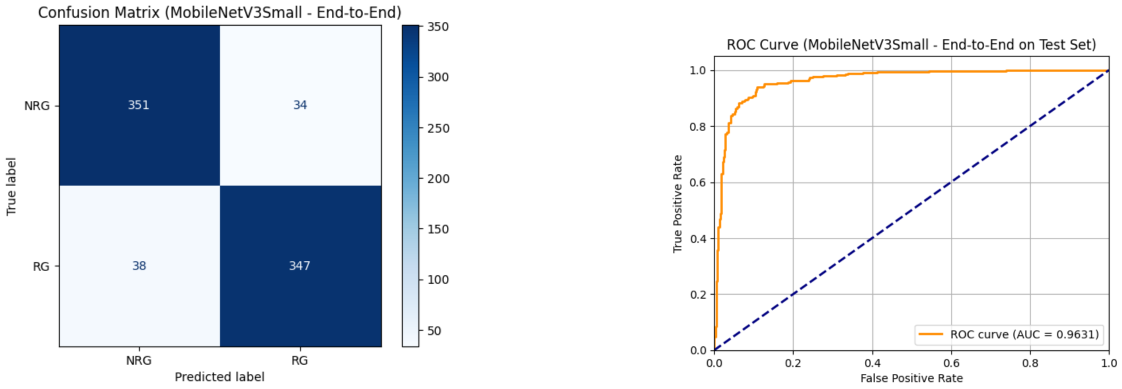
Figure 4.12: Confusion matrix and ROC curve (end-to-end ResNet50)



(a) Confusion matrix (MobileNetV2)

(b) ROC curve (MobileNetV2)

Figure 4.13: Confusion matrix and ROC curve (end-to-end MobileNetV2)



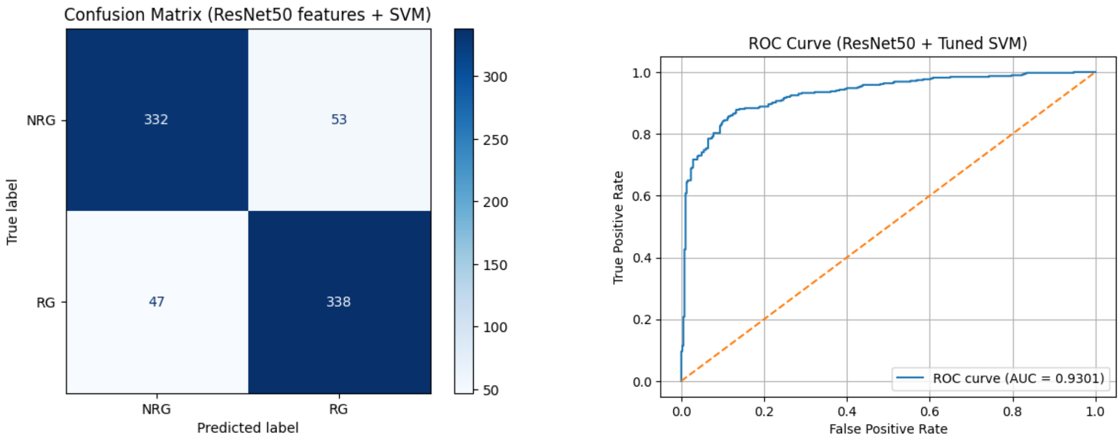
(a) Confusion matrix (MobileNetV3Small)

(b) ROC curve (MobileNetV3Small)

Figure 4.14: Confusion matrix and ROC curve (end-to-end MobileNetV3Small)

4.4 Hybrid approach

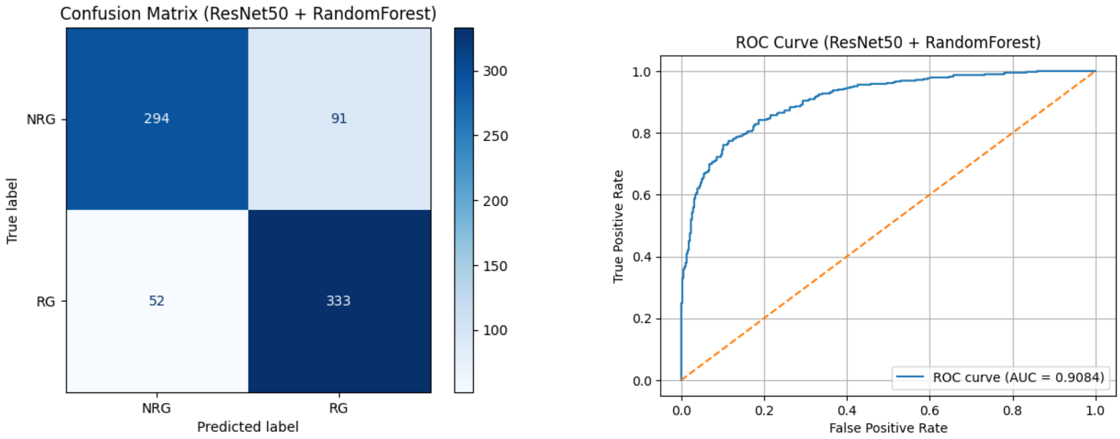
In the fourth group of experiments, we combined deep learning networks with traditional machine learning models. We used a pre-trained ResNet50 network as a feature extractor, inputting the extracted 2048 dimension feature vectors into two traditional machine learning classifiers: Support Vector Machine (SVM) and Random Forest for glaucoma classification. The hyperparameter search range for SVM included C uniformly distributed within 0.1-10, gamma within 0.0001-0.1, with fixed RBF kernel function. The hyperparameters for Random Forest included n estimators selected from [100, 200, 300], max depth varying among [None, 10, 20], min samples split from [2, 5, 10], and min samples leaf varying among [1, 2, 4]. Both algorithms used RandomizedSearchCV for hyperparameter optimization, with 20 search iterations, 3-fold cross-validation, and AUC-ROC as the evaluation metric.
As can be seen from the table 4.1, based on feature extraction using ResNet50, SVM achieved 87.01% accuracy, outperforming Random Forest's 81.43%. This indicates that SVM's kernel function mechanism can better handle the complex 2048-dimensional feature representations extracted by ResNet50. From the confusion matrix in the figure 4.15, we can see the ResNet50 + SVM model misclassified 53 NRG samples as RG and 47 RG samples as NRG, indicating that its performance on the two categories is relatively balanced. The confusion matrix in the figure 4.16 shows that the ResNet50 + Random Forest model misclassified 91 NRG samples as RG and 52 RG samples as NRG. The number of misclassifications is significantly higher than that of the ResNet50 + SVM model. It also means the model is unbalanced on the two categories, and its ability to classify the NRG samples is relatively weak. The AUC-ROC value of ResNet50 + SVM is 0.9301, and the AUC-ROC value of ResNet50 + Random Forest model is 0.9084, indicating that both of them are reliable in this task. However, compared with the results of the third group of experiments, end-to-end learning still performs better. This may be because when used as a feature extractor, ResNet50's weights are fixed and cannot be adjusted for the specific characteristics of medical images, further proving the importance of feature space adaptability.



(a) Confusion matrix (ResNet50+SVM)

(b) ROC curve (ResNet50+SVM)

Figure 4.15: Confusion matrix and ROC curve (ResNet50+SVM)



(a) Confusion matrix (ResNet50+RF)

(b) ROC curve (ResNet50+RF)

Figure 4.16: Confusion matrix and ROC curve (ResNet50+RandomForest)

4.5 Gradient-weighted Class Activation Mapping

As mentioned earlier, people might not trust deep learning models due to the lack of interpretability. To reduce the distrust and test our assumption that the two-stage transfer learning model cannot adapt to the unique features of this specific medical dataset, we used Gradient-weighted Class Activation Mapping (GRAD-CAM) to visualize the models' interest areas when they made predictions.

4.5.1 True positive examples

From figures 4.17 to 4.23, we can see that all models mainly focus on the optic disc area when making predictions, which is consistent with real medical diagnoses, indicating that the features used by the models are reasonable.

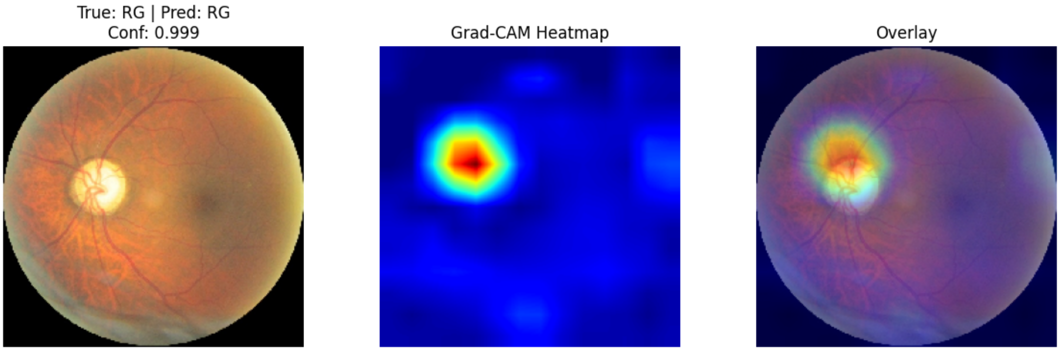


Figure 4.17: Classic CNN True positive

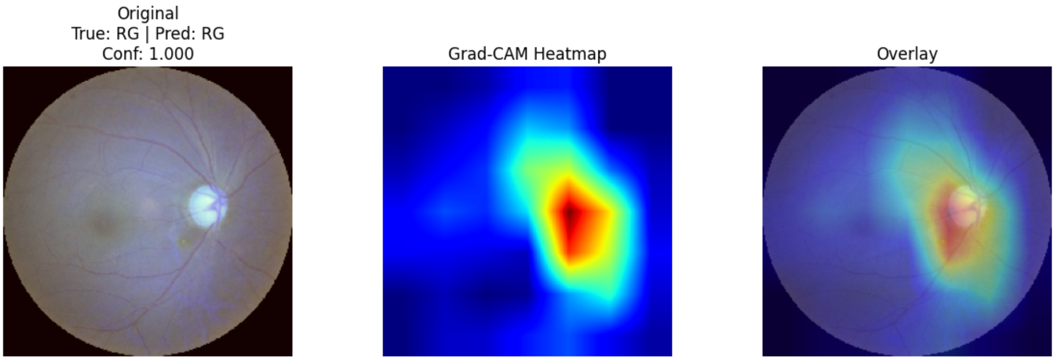


Figure 4.18: Two-stage ResNet50 True positive

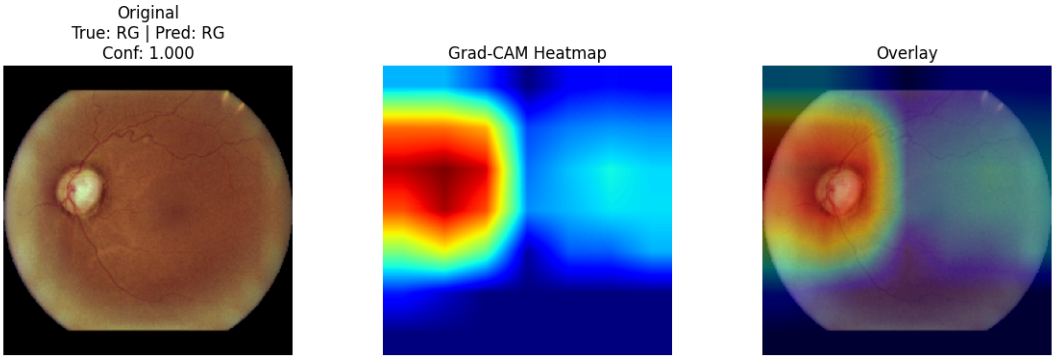


Figure 4.19: Two-stage MobileNetV2 True positive

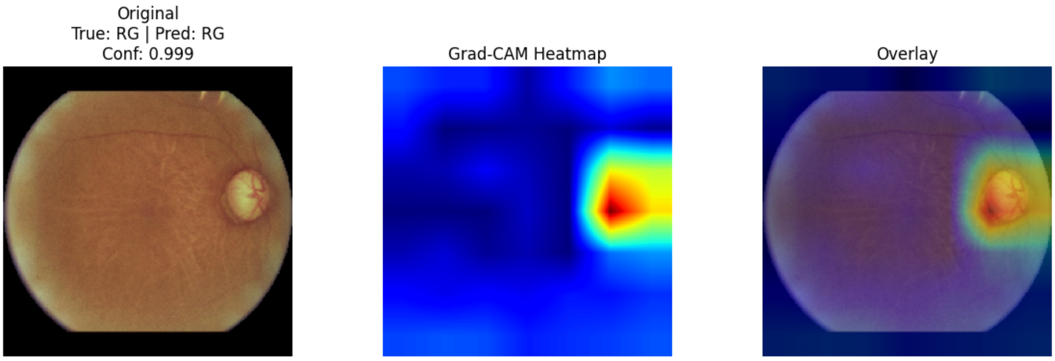


Figure 4.20: Two-stage MobileNetV3Small True positive

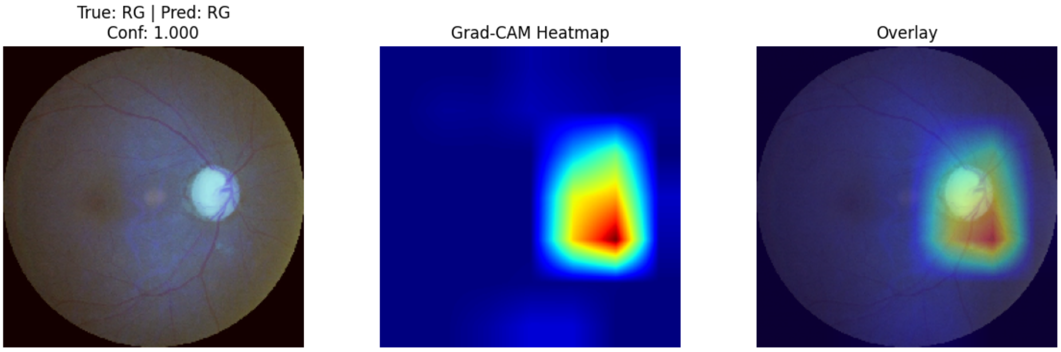


Figure 4.21: End-to-End ResNet50 True positive

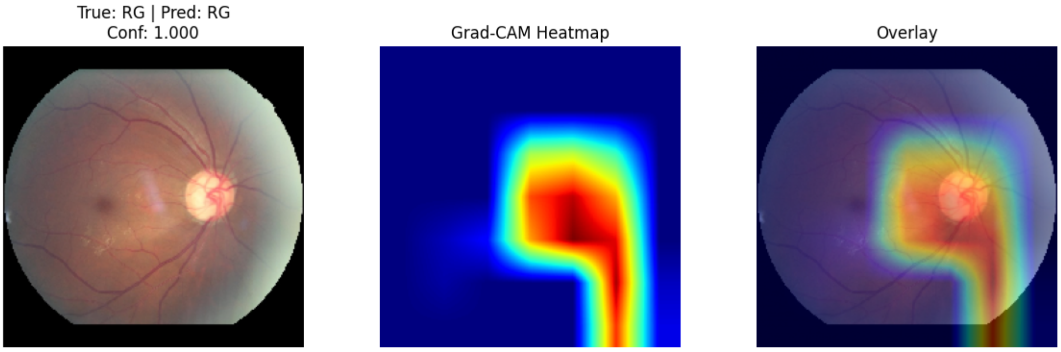


Figure 4.22: End-to-End MobileNetV2 True positive

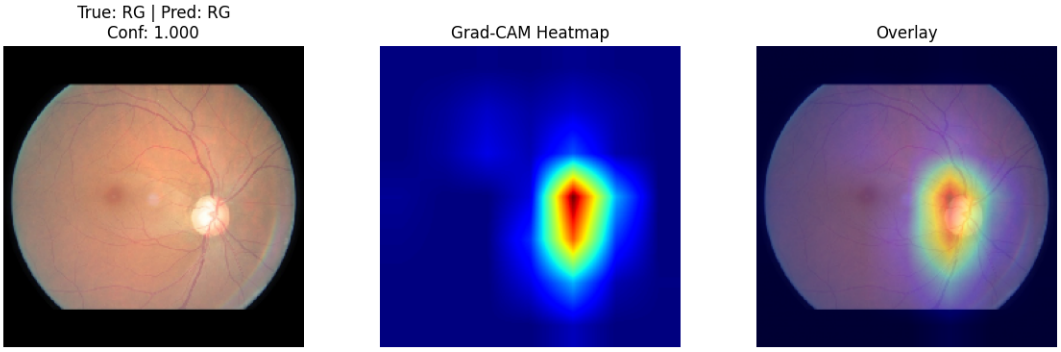


Figure 4.23: End-to-End MobileNetV3Small True positive

4.5.2 True negative examples

However, Figures 1 to 7 show that some models failed to detect the optic disc when making negative class predictions. Models like Classic CNN and End-to-End ResNet50 focus on the edges of the image, which can lead to incorrect predictions. To solve this problem, future research could try to do segmentation before inputting data into the predictive model.

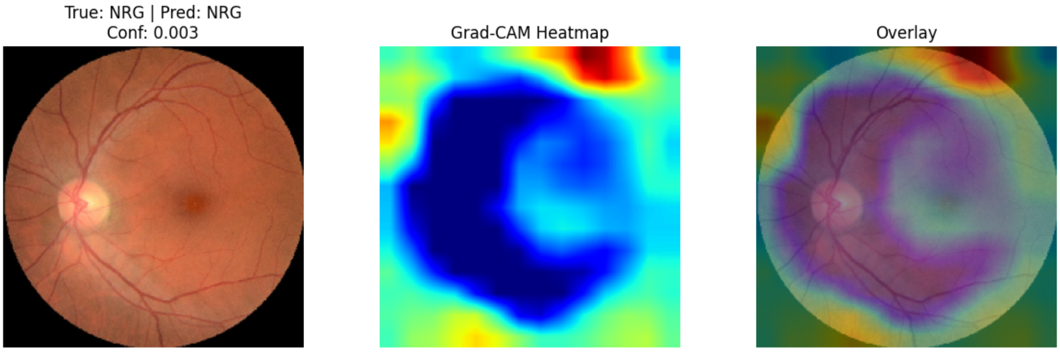


Figure 4.24: Classic CNN True negative

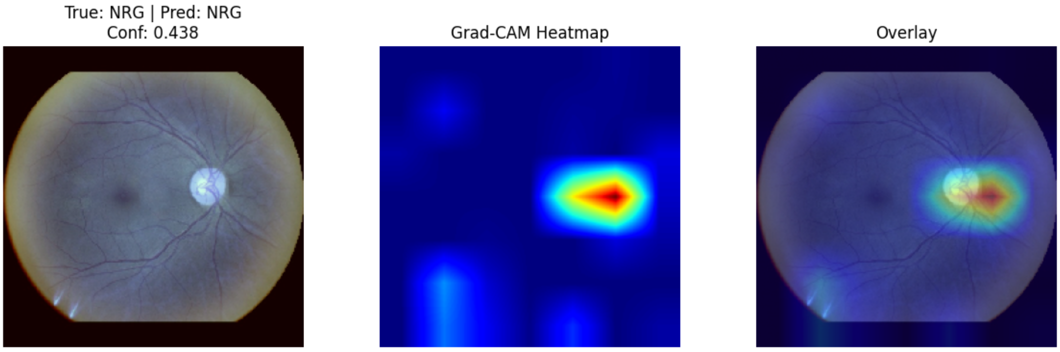


Figure 4.25: Two-stage ResNet50 True negative

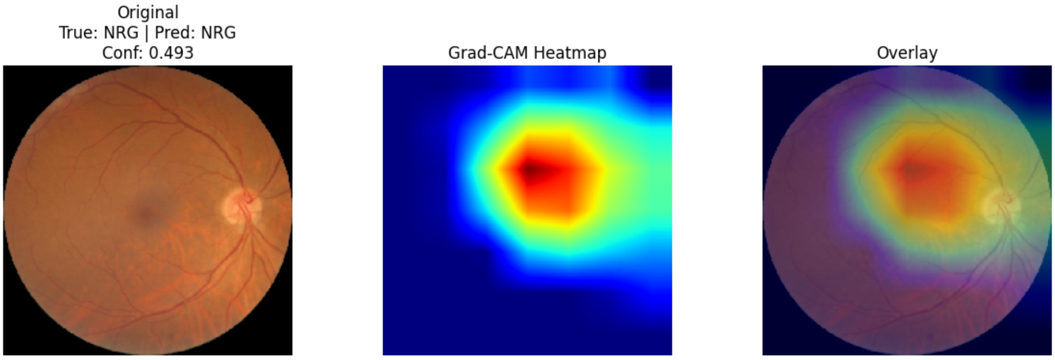


Figure 4.26: Two-stage MobileNetV2 True negative

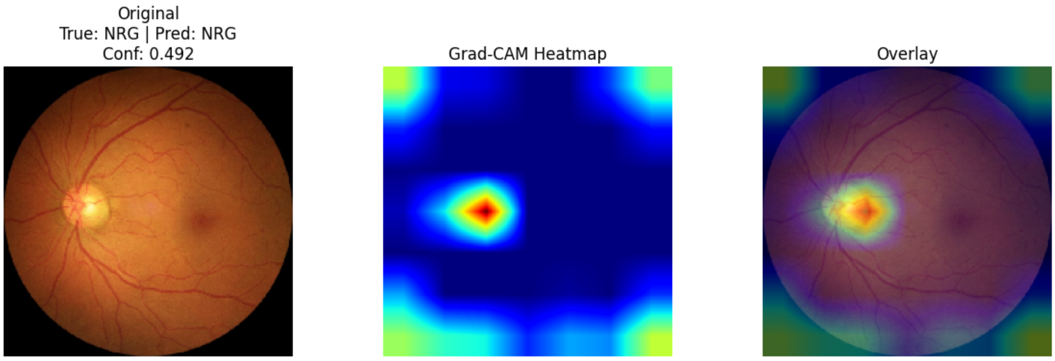


Figure 4.27: Two-stage MobileNetV3Small True negative

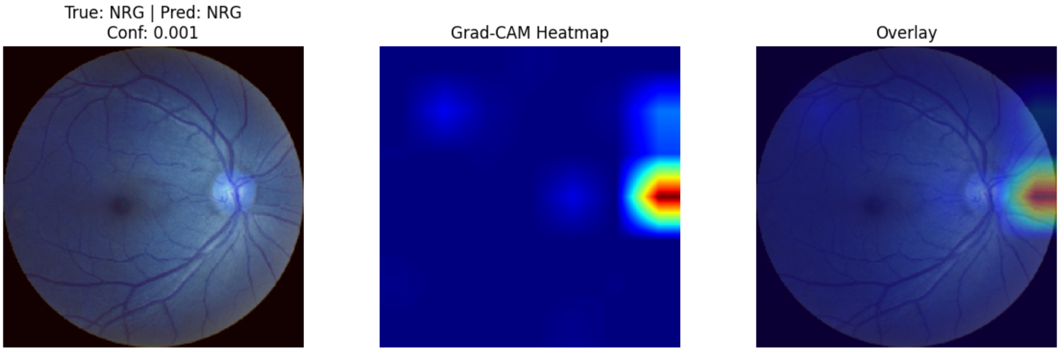


Figure 4.28: End-to-End ResNet50 True negative

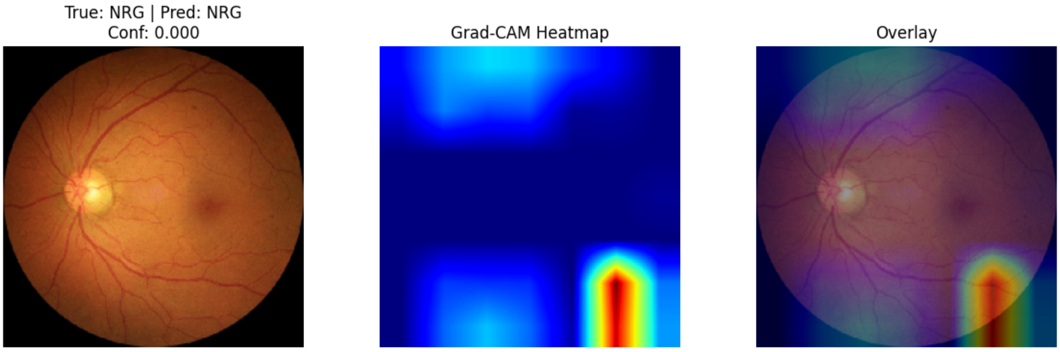


Figure 4.29: End-to-End MobileNetV2 True negative

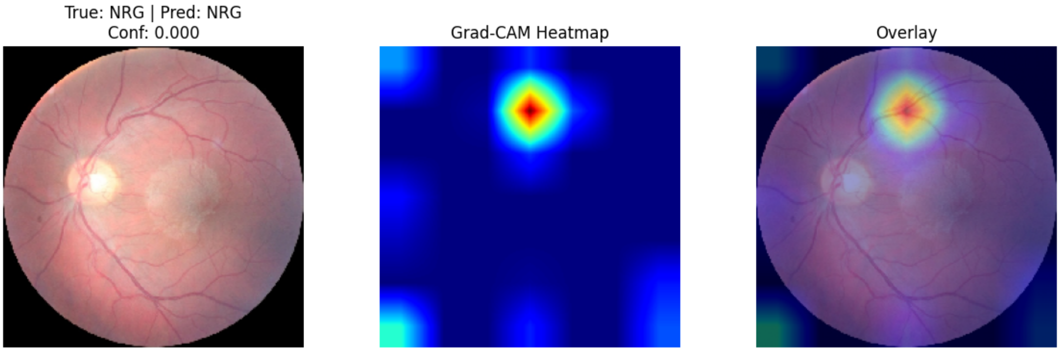


Figure 4.30: End-to-End MobileNetV3Small True negative

Chapter 5

Discussion and Analysis

5.1 Analysis of results

The experimental results show a clear performance difference. The classic stacked CNN, used as our baseline model and trained from scratch, achieved an accuracy of 83.12% and an AUC-ROC of 0.9143 on the test set. It means that even without pre-trained weights or complex architectures, a simple CNN model can effectively complete the glaucoma classification task on this dataset. The two-stage transfer learning approach showed improvements over this baseline. The ResNet50-based model with ImageNet pre-trained weights achieved the highest accuracy of 89.22%. However, the improvement was limited compared to our expectations. This limitation is because the first stage freezes the lower-layer weights of the model, limiting its ability to learn task-specific features from medical images, such as the optic disc morphology and nerve fiber layer thickness in this dataset. The end-to-end transfer learning strategy improves all three models significantly. The model based on ResNet50 achieved an accuracy of 92.34%, an AUC-ROC of 0.9756, and a recall of 96.10%. This improvement shows the benefits of end-to-end training in medical image analysis. By allowing the entire network to adapt to the target task, low-layer weights can be optimized based on unique textures and color patterns in medical images, enabling deeper and task-specific feature learning. In addition, lightweight models also showed excellent performance with end-to-end training. The accuracy of MobileNetV3-Small increased from 84.81% to 90.65%, and MobileNetV2 achieved 92.47%, showing the potential of lightweight architectures. These results offer practical solutions for deploying medical image classification systems with constrained resources. In the fourth experiment, we used ResNet50 as a feature extractor and combined it with traditional machine learning algorithms. The model combined with SVM achieved an accuracy of 87.01%, while the random forest model only achieved 81.43%. Both of them were lower than the end-to-end trained ResNet50. It further proves the importance of feature space adaptability.

Overall, our comparative experiments validate the effectiveness of pre-trained models with end-to-end strategies for glaucoma classification on the EyePACS-AIROGS-light-V2 dataset. We also proved the potential of lightweight models such as MobileNetV2 and MobileNetV3-Small. The performance difference between end-to-end training and two-stage training indicates that adapting the entire network to the dataset might have more advantages when dealing with medical image data that differs from natural images.

5.2 Limitations

Although we obtained these results from our experiments, there are still some limitations. First, all our experiments were conducted on a standardized dataset (EyePACS-AIROGS-light-V2).

The distributions of RG and NRG in the training, validation, and test sets are balanced, and the images are preprocessed. This may reduce the complexity of the original dataset. And the image data in the clinic may have class imbalance and varying image quality, both of which can impact the model's real-world performance.

Additionally, since the dataset was already split, we did not evaluate the robustness of the models under different data split strategies or cross-validation methods. Furthermore, this study only focused on the binary classification of RG and NRG using fundus images. The findings may not generalize to predictive tasks for other diseases, as those images might have different features and require different methods.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Our study explored predictive data analysis techniques in digital health applications through a systematic literature review and then evaluated the performance of multiple CNN based methods on glaucoma classification tasks using the EyePACS-AIROGS-light-V2 dataset through four comparative experiments.

We first used ASReview to screen 7,323 articles from four major databases, and considering the impact on real medical systems, we selected 249 articles for a systematic literature review to analyze the usage patterns of predictive analysis techniques in digital health applications. We found a correlation between the data format and data analysis techniques. Structured numerical data is the most common category, mainly using traditional machine learning methods such as SVM and Random Forest. Image data analysis heavily relies on CNN architectures and is often enhanced through transfer learning strategies to deal with limited data in medical applications. In audio data processing, traditional feature engineering methods are still widely used, but there is also a trend toward converting signals into spectrograms and then using CNN to analyze them. Research using video data is relatively limited because it usually requires complex preprocessing procedures and relies on computer vision tools with higher computational resource requirements. However, with the spread of smart devices, video collection has become more convenient, and data analysis using video data will have greater potential. We found the important role of CNN architectures in medical image analysis in our literature review, so we used the EyePACS-AIROGS-light-V2 glaucoma dataset to conduct four comparative experiments to evaluate the effectiveness of different CNN variants and training strategies. We found that classic stacked CNNs already have efficient baseline performance (accuracy of 86.36%, AUC-ROC of 0.9377), indicating that even simple CNN architectures can effectively handle medical image classification tasks. Two-stage transfer learning models showed improvement, with ResNet50 achieving 87.66% accuracy, but due to frozen lower-layer weights, they cannot learn specific features of medical images, so the improvement was limited. End-to-end transfer learning strategies led to significant performance improvements for all models. ResNet50 achieved 91.95% accuracy, 94.55% recall, and 0.9704 AUC-ROC. End-to-end transfer learning also made lightweight models perform better, with MobileNetV3-Small achieving 91.69% accuracy and MobileNetV2 achieving 91.30% accuracy, showing their potential in medical applications with limited resources. When ResNet50 was used as a feature extractor combined with traditional classifiers, the model with SVM classifier achieved 87.01% accuracy, further proving the superiority of end-to-end trained CNN models in medical image analysis.

Overall, this study provides a reference for predictive data analysis technique selection in digital health applications through a systematic literature review and specifically analyzes the performance of multiple CNN variations in medical image data analysis through comparative experiments.

6.2 Future work

As mentioned earlier, although this study has achieved some results on the EyePACS-AIROGS-light-V2 dataset, there are still many limitations, and future work can improve these areas. First, since EyePACS-AIROGS-light-V2 is a standardized and balanced dataset extracted from the Rotterdam EyePACS AIROGS, the numbers of RG and NRG samples in the training, validation, and test sets are all equal. It is different from real clinical data. Future research could try to evaluate model performance using non-standardized images and class-imbalanced data in actual medical scenarios.

Additionally, this study only designed experiments for CNN based models on image format datasets. Future work could design experiments for other data formats and corresponding data analysis techniques to set benchmarks. For example, exploring the actual performance of RNN models in processing audio format medical data, standardizing the framework for medical video data collection and preprocessing, etc. Furthermore, this study only focused on data analysis techniques related to predictive analysis in digital health applications, mainly used for disease diagnosis and classification. There are many other data analysis techniques for different purposes that need further research exploration. In summary, due to the complexity of medical data in form and content and the characteristic that the medical industry is closely related to people's quality of life, there is still a lot of chance to explore data analysis in digital health applications.

References

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D. and Al-Muhanna, F. A. (2023), ‘A review of the role of artificial intelligence in healthcare’, *Journal of Personalized Medicine* **13**(6), 951.
URL: *https://doi.org/10.3390/jpm13060951*

Bajwa, J., Munir, U., Nori, A. and Williams, B. (2021), ‘Artificial intelligence in healthcare: transforming the practice of medicine’, *Future Healthcare Journal* **8**(2), e188–e194.
URL: *https://doi.org/10.7861/fhj.2021-0095*

Castillo-Escario, Y., Werthen-Brabants, L., Groenendaal, W., Deschrijver, D. and Jané, R. (2022), Convolutional neural networks for apnea detection from smartphone audio signals: Effect of window size, *in* ‘2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)’, Scottish Event Campus, Glasgow, UK, pp. 666–669.

Chamberlain, D., Kodgule, R., Ganelin, D., Miglani, V. and Fletcher, R. R. (2016), Application of semi-supervised deep learning to lung sound analysis, *in* ‘2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)’, IEEE, Orlando, FL, USA, pp. 804–807.

Chia, A. A., Lum, S., Boo, M., Tan, R., Nair, B. B. and Chen, J.-M. (2024), Transfer learning for dysphagia detection, *in* ‘TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON)’, IEEE, Malaysia, pp. 619–623.

Choi, Y., Choi, H., Lee, H., Lee, S. and Lee, H. (2022), ‘Lightweight skip connections with efficient feature stacking for respiratory sound classification’, *IEEE Access* **10**, 53027–53042.

Fazeli, S., Levine, L., Beikzadeh, M., Mirzasoleiman, B., Zadeh, B., Peris, T. and Sarrafzadeh, M. (2022), Passive monitoring of physiological precursors of stress leveraging smartwatch data, *in* ‘2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)’, IEEE, pp. 2893–2899.

Girmaw, D. W. and Taye, G. B. (2025), ‘Mobilenetv2 model for detecting and grading diabetic foot ulcer’, *Discover Applied Sciences* **7**, 268. Open access article under CC BY-NC-ND 4.0 license.
URL: *https://doi.org/10.1007/s42452-025-06745-4*

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep learning*, MIT press.

Guarín, D. L., Wong, J. K., McFarland, N. R., Ramirez-Zamora, A. and Vaillancourt, D. E. (2024), ‘What the trained eye cannot see: Quantitative kinematics and machine learning detect movement deficits in early-stage parkinson’s disease from videos’, *Parkinsonism & Related Disorders* **127**, 107104.

Guo, C., Ashrafian, H., Ghafur, S., Fontana, G., Gardner, C. and Prime, M. (2020), 'Challenges for the evaluation of digital health solutions—a call for innovative evidence generation approaches', *npj Digital Medicine* **3**(1), 110.

Güven, M., Hardalaç, F., Özışık, K. and Tuna, F. (2021), 'Heart diseases diagnose via mobile application', *Applied Sciences* **11**(5), 2430.
URL: *https://doi.org/10.3390/app11052430*

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V. et al. (2019), Searching for mobilenetv3, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 1314–1323.

Hu, H.-C., Chang, S.-Y., Wang, C.-H., Li, K.-J., Cho, H.-Y., Chen, Y.-T., Lu, C.-J., Tsai, T.-P. and Lee, O. K.-S. (2021), 'Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study', *Journal of Medical Internet Research* **23**(6), e25247.
URL: *https://www.jmir.org/2021/6/e25247*

Hwang, D.-K., Hsu, C.-C., Chang, K.-J., Chao, D., Sun, C.-H., Jheng, Y.-C., Yarmishyn, A. A., Wu, J.-C., Tsai, C.-Y., Wang, M.-L. et al. (2019), 'Artificial intelligence-based decision-making for age-related macular degeneration', *Theranostics* **9**(1), 232–245. Open access article under CC BY-NC license.
URL: *http://www.thno.org*

Jahin, S., Moniruzzaman, M., Alvee, F. M., Haque, I. U. and Kalpoma, K. A. (2022), A modern approach to ai assistant for heart disease detection by heart sound through created e-stethoscope, *in* '2022 25th International Conference on Computer and Information Technology (ICCIT)', IEEE, Cox's Bazar, Bangladesh, pp. 669–674.

Kashani, A. B., Baraeinejad, B. and Fakharzadeh, M. (2022), A new atrial fibrillation detection system with noise cancellation and signal annotation, *in* '2022 30th International Conference on Electrical Engineering (ICEE)', IEEE, Tehran, Iran, pp. 256–261.

Kiefer, R. (2024), 'EyePACS-AIROGS-light-V2'.
URL: *https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2*

Liu, S., Dong, J., Dong, J., Zhao, Y. and Wang, Y. (2023), Deep learning based fall detection using smartwatches for healthcare applications, *in* '2023 22nd IEEE International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)', IEEE, pp. 119–125.

Mathews, S. C., McShea, M. J., Hanley, C. L., Ravitz, A., Labrique, A. B. and Cohen, A. B. (2019), 'Digital health: a path to validation', *npj Digital Medicine* **2**(1), 38.

Mayr, W., Triantafyllopoulos, A., Batliner, A., Schuller, B. W. and Berghaus, T. M. (2025), 'Assessing the clinical and functional status of copd patients using speech analysis during and after exacerbation', *International Journal of Chronic Obstructive Pulmonary Disease* pp. 137–147.

Mejía, C. R., Flores, L. H., Rodríguez-Alarcón, J., Arias-Chávez, D., Valverde-Villacorta, L., Taype-Rondan, A., Huamán-Mamani, Y., Ponce, J., Dominguez-Vergara, J. and Aguilar-Linares, J. (2022), 'Remote evaluation of parkinson's disease using a conventional webcam and artificial intelligence', *Journal of Parkinson's Disease* **12**(2), 597–605.

Motin, M. A., Pah, N. D., Raghav, S. and Kumar, D. K. (2022), 'Parkinson's disease detection using smartphone recorded phonemes in real world conditions', *IEEE Access* **10**, 97600–97609.

Narváez, P., Vera, K., Bedoya, N. and Percybrooks, W. S. (2017), Classification of heart sounds using linear prediction coefficients and mel-frequency cepstral coefficients as acoustic features, *in* '2017 IEEE Colombian Conference on Communications and Computing (COLCOM)', IEEE, pp. 1–6.

Ngeh, C. J., Ma, C., Ho, T. K.-W., Wang, Y. and Raiti, J. (2020), Deep learning on edge device for early prescreening of skin cancers in rural communities, *in* '2020 IEEE Global Humanitarian Technology Conference (GHTC)', IEEE, pp. 1–5.

Organization, W. H. (2023), *Classification of digital interventions, services and applications in health: A shared language to describe the uses of digital technology for health*, World Health Organization.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. et al. (2021), 'The prisma 2020 statement: an updated guideline for reporting systematic reviews', *BMJ* **372**, n71.

Pan, J. and Nan, F. (2024), Quantifying and removing free-living uncertainty for effective parkinson's disease diagnosis using smart watch, *in* '2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)', IEEE, pp. 5976–5983.

Pangti, R., Mathur, J., Chouhan, V., Kumar, S., Rajput, L., Shah, S., Gupta, A., Dixit, A., Dholakia, D., Gupta, S., Gupta, S., George, M., Sharma, V. and Gupta, S. (2021), 'A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases', *Journal of the European Academy of Dermatology and Venereology* **35**(2), 536–545.

Park, B., Kim, M., Jung, D., Lee, D., Kim, J. and Mun, K.-R. (2024), Classification of abnormal gaits with machine learning algorithms using sensor-inherited insoles, *in* '2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, p. TBD.

Park, J. Y., Chun, M. H., Kim, H., Kim, W. J., Kim, B. J. and Lee, H. S. (2023), 'Prediction of motor unified parkinson's disease rating scale scores in patients with parkinson's disease using surface electromyography', *Heliyon* **9**(11), e21023.

Peyroteo, M., Ferreira, I. A., Elvas, L. B., Ferreira, J. C. and Lapão, L. V. (2021), 'Remote monitoring systems for patients with chronic diseases in primary health care: Systematic review', *JMIR mHealth and uHealth* **9**(12), e28285.

Pourazad, M. T., Shojaei-Hashemi, A., Nasiopoulos, P., Azimi, M., Mak, M., Grace, J., Jung, D. and Bains, T. (2020), A non-intrusive deep learning based fall detection scheme using video cameras, *in* '2020 International Conference on Information Networking (ICOIN)', IEEE, pp. 443–446.

Price, W. N. and Cohen, I. G. (2019), 'Privacy in the age of medical big data', *Nature Medicine* **25**(1), 37–43.

Radouani, L., Lagdali, S. and Rziza, M. (2021*a*), 'Detection of voice impairment for parkinson's disease using machine learning tools', *2020 10th International Symposium on Signal, Image, Video and Communications (ISIVC)* pp. 1–6.

Radouani, L., Lagdali, S. and Rziza, M. (2021*b*), Detection of voice impairment for parkinson's disease using machine learning tools, *in* '2021 10th International Symposium on Signal, Image, Video and Communications (ISIVC)', IEEE, pp. 1–6.

Randazzo, V., Buccellato, P., Ferretti, J., Delrio, F. and Pasero, E. (2024), Pulsecg: A cuffless non-invasive blood pressure monitoring device through neural network analysis of ecg and ppg signals, *in* '2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)', IEEE, pp. 1030–1035.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018), Mobilenetv2: Inverted residuals and linear bottlenecks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4510–4519.

Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Ettlin, D. A. et al. (2023), 'Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review', *PLOS Digital Health* **2**(10), e0000347.

Soguero-Ruiz, C., Hindberg, K., Rojo-Álvarez, J. L., Skrøvseth, S. O., Godtliebsen, F., Mortensen, K., Revhaug, A., Lindsetmo, R.-O., Augestad, K. M. and Jenssen, R. (2016), 'Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records', *IEEE Journal of Biomedical and Health Informatics* **20**(5), 1404–1415.

Steen, J., Kiefer, R., Ardali, M. R., Abid, M. and Amjadian, E. (2023), 'Standardized and open-access glaucoma dataset for artificial intelligence applications', *Investigative Ophthalmology & Visual Science* **64**(8), 384–384.

Tan, S. Y., Sumner, J., Wang, Y. and Yip, A. W. (2024), 'A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes', *npj Digital Medicine* **7**(1), 192.

Tariq, S., Akhtar, N., Afzal, H., Khalid, S., Mufti, M. R., Hussain, S., Habib, A. and Ahmad, G. (2019), 'A novel co-training-based approach for the classification of mental illnesses using social media posts', *IEEE Access* **7**, 166165–166172.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L. and Oberski, D. L. (2021), 'An open source machine learning framework for efficient and transparent systematic reviews', *Nature Machine Intelligence* **3**(2), 125–133.
URL: *https://doi.org/10.1038/s42256-020-00287-7*

Vatanparvar, K., Nathan, V., Nemati, E., Rahman, M. M., McCaffrey, D., Kuang, J. and Gao, J. (2021), Speechspiro: Lung function assessment from speech pattern as an alternative to spirometry for mobile health tracking, *in* '2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)', IEEE, pp. 7237–7243.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S. and Jantana, P. (2021), 'Automatic classification and detection of oral cancer in photographic images using deep learning algorithms', *Journal of Oral Pathology & Medicine* **50**(9), 911–918.

Wosik, J., Fudim, M., Cameron, B., Gellad, Z. F., Cho, A., Phinney, D., Curtis, S., Roman, M., Poon, E. G., Ferranti, J. et al. (2020), 'Telehealth transformation: COVID-19 and the rise of virtual care', *Journal of the American Medical Informatics Association* **27**(6), 957–962.

Xiang, Y., Li, S. and Zhang, P. (2022), 'An exploration in remote blood pressure management: Application of daily routine pattern based on mobile data in health management', *Fundamental Research* **2**(2), 154–165. Available online 16 November 2021.
URL: *https://doi.org/10.1016/j.fmre.2021.11.006*

Yang, Y., Liu, P., Sun, Y., Yu, N., Wu, J. and Han, J. (2021), A video-based method to classify abnormal gait for remote screening of parkinson's disease, *in* 'Proceedings of the 40th Chinese Control Conference', IEEE, pp. 3357–3362. Accessed via IEEE Xplore.
URL: *https://ieeexplore.ieee.org/document/9530162*

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014), How transferable are features in deep neural networks?, *in* 'Advances in neural information processing systems', Vol. 27, pp. 3320–3328.

Zhang, L., Qu, Y., Jin, B., Jing, L., Gao, Z. and Liang, Z. (2020), 'An intelligent mobile-enabled system for diagnosing parkinson disease: Development and validation of a speech impairment detection system', *JMIR Medical Informatics* **8**(9), e18689.
URL: *http://medinform.jmir.org/2020/9/e18689/*

Zhang, P., Zhang, X., Teng, M., Li, L., Liu, X., Feng, J., Wang, W., Wang, X. and Luo, X. (2024), 'Leather-based shoe soles for real-time gait recognition and automatic remote assistance using machine learning', *ACS Applied Materials Interfaces* **16**(45), 62803–62816.

Zheng, Q., Huang, Q., Chen, Y., Cui, K., You, C., Chen, Y., Xu, L., Ma, L., Liu, N., Xie, Y. et al. (2022), 'Severe aortic stenosis detection by deep learning applied to echocardiography', *JACC: Cardiovascular Imaging* **15**(3), 438–448.