Deep Learning for Retinal Disease Classification: Comparative Analysis of RFMiD, APTOS, and Combined Datasets

Etana Disasa
Department of Data Science
College of Computer and Information Sciences
Regis University
3333 Regis Boulevard, Denver, CO 80221
edisasa@regis.edu

June 2025

Abstract

Early detection of retinal diseases is essential for preventing vision loss, yet access to timely ophthalmological diagnosis is limited, particularly in low-resource settings. This practicum project explores the use of deep learning techniques to automate the classification of retinal diseases using fundus images. Transfer learning was employed using MobileNetV2, a lightweight convolutional neural network architecture, to develop and compare three image classification models trained on different datasets: RFMiD, APTOS 2019 Blindness Detection, and a combined dataset merging the two. Each model underwent data augmentation, preprocessing, and rigorous evaluation. Performance metrics included accuracy, loss, learning curves, and confusion matrices. The best-performing model was deployed via a Streamlit web application, enabling real-time predictions. The project highlights strengths and limitations of training on individual versus combined datasets and discusses implications for AI-driven retinal disease screening.

1 Introduction

Retinal diseases, including diabetic retinopathy, are leading causes of vision impairment and blindness worldwide. Such conditions often progress silently and require specialized eye examinations using fundus imaging for diagnosis. Access to these diagnostic services is limited in many regions, especially in low- and middle-income countries, due to a shortage of specialists and equipment costs. Artificial intelligence, particularly deep learning, offers promising solutions for automating retinal disease detection from fundus images (Gulshan et al.,). Convolutional neural networks (CNNs) have demonstrated near-clinician level performance in image classification tasks. This study utilizes MobileNetV2, optimized for efficiency and suitable for edge deployment, to train models on RFMiD and APTOS datasets individually and combined, aiming to build accurate and deployable retinal disease classifiers.

2 Problem Statement

The central question is whether a deep learning model can accurately classify retinal fundus images into binary diabetic retinopathy categories using data from multiple sources. Ensuring label consistency across heterogeneous datasets such as RFMiD and APTOS is essential to improve reliability and avoid misclassification. The study evaluates model performance on combined versus individual datasets.

3 Related Work

Deep learning has revolutionized diabetic retinopathy detection, with early work by Gulshan et al. (Gulshan et al.,) developing a CNN achieving high accuracy for retinal fundus image screening. Subsequently, datasets like APTOS and RFMiD have become benchmarks for DR classification tasks. Takahashi et al. (Takahashi, Tampo, Arai, Inoue, & Kawashima,) applied deep learning for DR staging, highlighting clinical relevance. Rajalakshmi et al. (Rajalakshmi, Subashini, Anjana, & Mohan,) explored automated DR screening using digital fundus images. Multi-task learning approaches by Lin et al. (Lin, Yang, Hsieh, Lin, & Shieh,) have improved classification accuracy by leveraging shared representations across retinal diseases. Despite these advances, challenges remain in harmonizing diverse datasets to improve model generalization.

4 Methodology

4.1 Data Sources and Preprocessing

Publicly available datasets RFMiD (Retinal Fundus Multi-Disease Image Dataset (RFMiD),) and APTOS 2019 (Asia Pacific Tele-Ophthalmology Society (APTOS),) were used. The RFMiD dataset included a binary diabetic retinopathy (DR) label, while APTOS contained five severity levels converted to binary (presence or absence of DR) for label consistency. Images were resized to 224×224 pixels and normalized to [0,1]. Data augmentation techniques included horizontal flipping, rotation (up to 15 degrees), zooming, and shifting. RGB channels were preserved to match MobileNetV2's pretrained input format.

4.2 Model Architecture

MobileNetV2 served as the backbone CNN due to its efficiency and accuracy balance. Early layers of the pretrained model were frozen, with fine-tuning applied to later layers. The architecture included global average pooling, batch normalization, dense, dropout layers, and a sigmoid output layer for binary classification. Callbacks such as early stopping and learning rate reduction were employed to prevent overfitting.

4.3 Training and Evaluation

Models were trained on RFMiD, APTOS, and a combined dataset separately, using binary cross-entropy loss. Each model trained for up to 20 epochs with batch size 32. Early stopping (patience 5) and learning rate reduction on plateau (factor 0.5, patience 3) were used. Performance was evaluated via accuracy, loss, learning curves, and confusion matrices.

4.4 Deployment Using Streamlit

To demonstrate practical applicability, a Streamlit web application was developed to deploy the best-performing model. The app enables users to upload retinal fundus images and receive real-time classification predictions. It integrates model loading, preprocessing, and inference within an interactive and user-friendly interface. This prototype illustrates the potential for telemedicine and point-of-care screening applications.

5 Results and Discussion

5.1 Model Performance

The three deep learning models developed in this study—trained on the RFMiD dataset, APTOS dataset, and the combined dataset—demonstrated distinct performance patterns during training and validation.

5.1.1 RFMiD Model

The RFMiD model began with a modest training accuracy of approximately 52.5% in Epoch 1 and achieved progressive improvements, reaching training accuracies above 89% in later epochs. Its validation accuracy peaked at around 84%, with validation losses steadily decreasing when improvements occurred. The learning curves indicated consistent learning, though occasional fluctuations suggested the model's sensitivity to class distributions and potential overfitting in certain epochs.

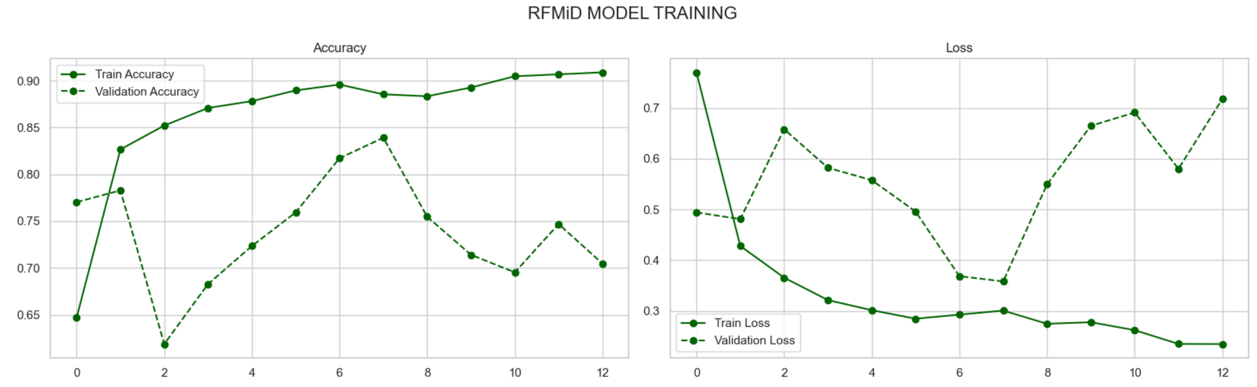

Figure 1: RFMiD Model Learning Curve

Table 1: RFMiD Model Training Summary (TA: Training Accuracy, TL: Training Loss, VA: Validation Accuracy, VL: Validation Loss)

Epoch	TA (%)	TL (%)	VA (%)	VL (%)	Notes
1	52.5	102.3	77.0	49.4	val_loss improved, model saved
2	81.7	43.4	78.3	48.1	val_loss improved, model saved
3	85.3	36.1	61.9	65.8	val_loss did not improve
4	86.9	32.2	68.3	58.2	val_loss did not improve
5	88.6	28.8	72.3	55.8	ReduceLROnPlateau triggered
6	89.9	27.1	75.9	49.6	val_loss did not improve
7	89.3	29.4	81.7	36.9	val_loss improved, model saved
8	89.7	27.4	83.9	35.8	val_loss improved, model saved
9	88.4	27.6	75.5	54.9	val_loss did not improve
10	89.8	26.8	71.4	66.5	val_loss did not improve
11	90.1	28.2	69.5	69.1	ReduceLROnPlateau triggered
12	91.0	23.1	74.7	58.0	val_loss did not improve
13	90.3	24.6	70.5	71.8	val_loss did not improve

5.1.2 APTOS Model

The APTOS model showed high initial training accuracy of nearly 80%, which quickly rose above 95% by mid-training. Its validation accuracy also improved, reaching as high as 96% in later epochs, with corresponding validation losses decreasing significantly to approximately 0.15. The model demonstrated robust generalization capabilities on the validation dataset, and learning rate reductions triggered by plateauing validation losses supported additional performance gains.

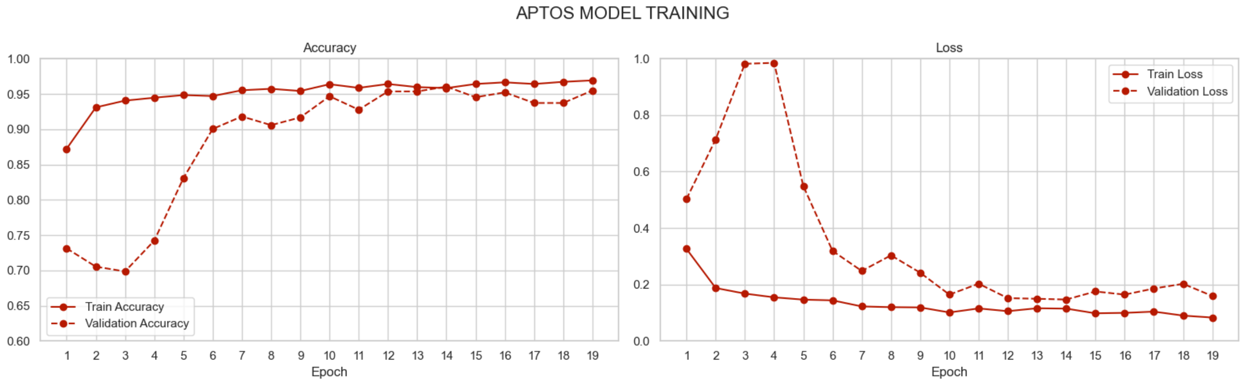

Figure 2: RFMiD Model Learning Curve

Table 2: RFMiD Model Training Summary (TA: Training Accuracy, TL: Training Loss, VA: Validation Accuracy, VL: Validation Loss)

Epoch	TA (%)	TL	VA (%)	VL	Notes
1	79.9	0.464	73.1	0.503	val_loss improved, model saved
2	91.7	0.212	70.5	0.714	
3	93.6	0.180	69.9	0.982	
4	94.9	0.137	74.2	0.985	ReduceLROnPlateau
5	95.4	0.133	83.1	0.549	
6	94.9	0.142	90.0	0.319	val_loss improved, model saved
7	95.2	0.131	91.8	0.248	val_loss improved, model saved
8	96.4	0.105	90.6	0.303	
9	94.9	0.143	91.7	0.242	val_loss improved, model saved
10	95.9	0.109	94.7	0.164	val_loss improved, model saved
11	95.9	0.124	92.8	0.203	
12	96.0	0.112	95.4	0.152	val_loss improved, model saved
13	96.1	0.115	95.4	0.150	val_loss improved, model saved
14	95.4	0.121	96.0	0.147	val_loss improved, model saved
15	96.5	0.092	94.5	0.175	
16	95.6	0.124	95.2	0.164	
17	96.7	0.096	93.7	0.186	ReduceLROnPlateau
18	97.2	0.080	93.7	0.203	
19	97.0	0.079	95.5	0.159	

5.1.3 Combined Model

The combined dataset model started with a training accuracy of about 73.6%, which rapidly increased to over 91% by Epoch 4. Validation accuracy peaked at approximately 88%, with an initial steep decrease in validation loss. However, from Epoch 4 onwards, fluctuations in validation accuracy and loss suggested potential challenges due to data heterogeneity across combined sources. Learning rate reductions mitigated this and maintain model stability.

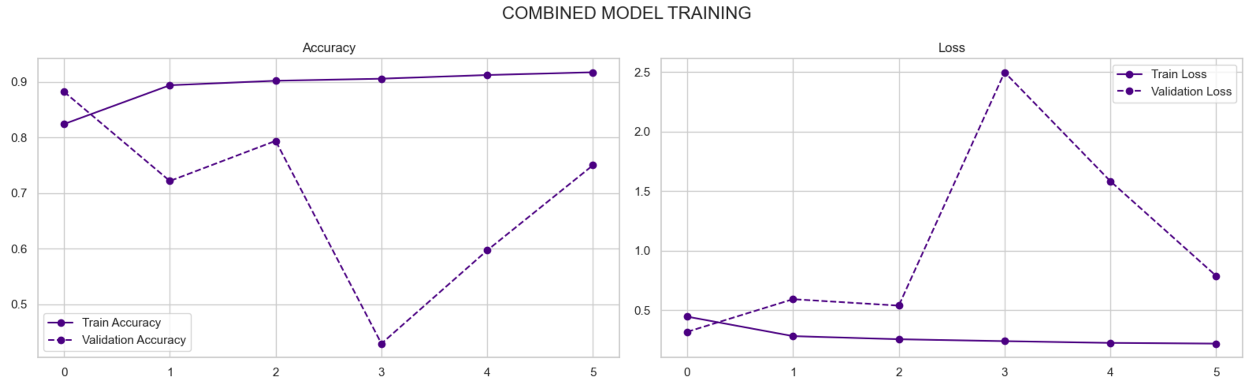

Figure 3: Enter Caption

Epoch	TA (%)	TS (%)	VA (%)	VL (%)	Notes
1	73.57	63.70	88.20	31.68	val_loss improved, saved best model
2	88.43	29.42	72.18	59.09	val_loss did not improve
3	89.95	25.72	79.39	53.69	val_loss did not improve
4	90.99	23.14	42.97	249.79	ReduceLROnPlateau triggered
5	91.39	22.37	59.72	158.26	val_loss did not improve
6	91.69	22.29	75.02	78.66	val_loss did not improve

Table 3: Combined Dataset Model Training Results. Ep=Epoch, TA=Training Accuracy, TS=Training Loss, VA=Validation Accuracy, VL=Validation Loss.

Overall, each model demonstrated effective learning, with clear improvements in both training and validation performance metrics across epochs. The use of callbacks, such as early stopping and learning rate reduction on plateau, contributed to model optimization and prevention of overfitting. The performance tables provided under each model section summarize these trends, illustrating epoch-wise accuracies, losses, and notes on model saving or learning rate adjustments.

These results provide a foundation for the subsequent comparative insights section, which analyzes the strengths and limitations of each model in relation to dataset characteristics, disease class diversity, and overall predictive utility.

5.2 Comparative Insights

Models trained on individual datasets outperformed the combined model, suggesting dataset heterogeneity—such as differences in imaging, labels, and patient demographics—impacts performance. Naïve combination without domain adaptation or harmonization reduced model generalization and increased overfitting risk.

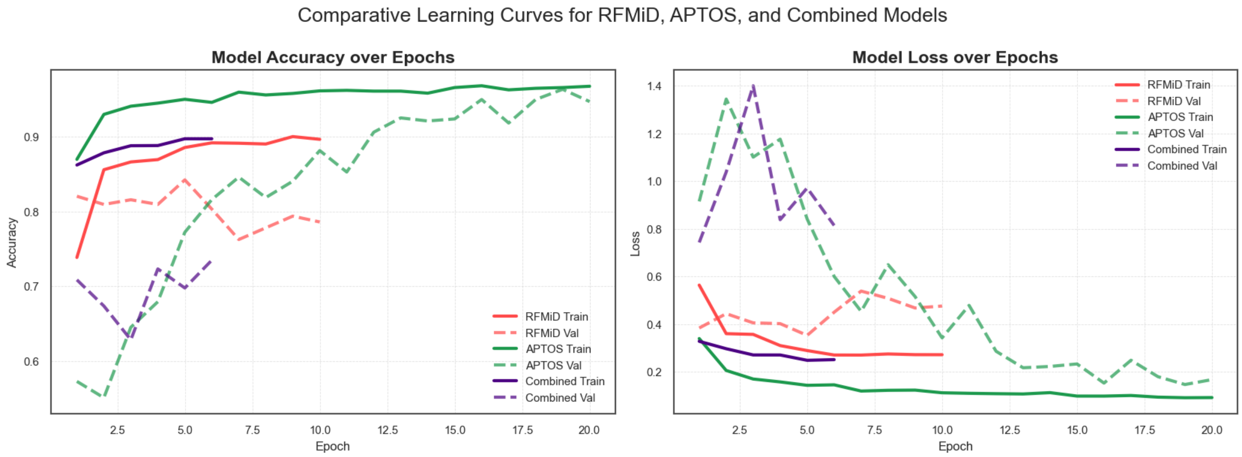

Figure 4: Enter Caption

Table 4: Comparative Summary of Model Performances

Metric	**RFMiD Model**	**APTOS Model**	**Combined Model**
Initial Training Accuracy	∼52.5%	∼79.9%	∼73.6%
Peak Training Accuracy	∼91.0%	∼97.2%	∼91.7%
Peak Validation Accuracy	∼84%	∼96%	∼88%
Lowest Validation Loss	∼0.35	∼0.14	∼0.31
Number of Epochs Trained	13	19	6
Learning Rate Reduction	Triggered twice	Triggered twice	Triggered once
Early Stopping Triggered	Not triggered	Not triggered	Stopped at Epoch 6

5.3 Implications for Multi-Dataset Training

Training models on combined datasets introduces challenges due to variations in imaging quality, labeling standards, and patient demographics across sources. To address these limitations, future studies should incorporate domain adaptation techniques to align feature distributions between datasets, ensuring better generalization. Additionally, implementing standardized preprocessing pipelines can harmonize image inputs, while multi-task learning frameworks can leverage shared representations across related disease categories to improve overall performance. Finally, ensemble approaches that integrate predictions from multiple specialized models may enhance robustness and mitigate the risks associated with dataset heterogeneity and label noise.

5.4 Clinical Relevance/Contribution

This study demonstrates that deep learning models, particularly MobileNetV2-based architectures trained on high-quality single datasets, can achieve high accuracy in detecting diabetic retinopathy from fundus images, supporting their potential use as screening tools in clinical settings. By systematically comparing models trained on RFMiD, APTOS, and combined datasets, this research contributes to knowledge by highlighting the impact of dataset heterogeneity on model performance and the challenges of multi-source data integration for medical AI applications.

Furthermore, the project provides practical insights into **model deployment workflows**, showcasing a working Streamlit application as a proof of concept for real-time retinal disease screening. This bridges the gap between model development and clinical utility, underlining the importance of standardized data preprocessing, harmonization techniques, and deployment pipelines for effective AI integration in ophthalmology.

Overall, this research extends current understanding of dataset selection strategies, transfer learning performance, and deployment considerations, thereby informing future AI development efforts targeting equitable, accessible, and scalable retinal disease screening solutions worldwide.

5.5 Future Work

This project successfully developed a Streamlit web application using the best-performing model—the APTOS-trained model—to enable real-time diabetic retinopathy predictions from uploaded fundus images. However, deployment revealed limitations related to image formats, resolutions, and varying image quality, which affected model input compatibility and prediction consistency.

Future development will focus on expanding the application to support multi-class disease classification for broader clinical utility, integrating explainable AI visualizations such as Grad-CAM heatmaps to improve model interpretability for clinicians, and establishing interoperability with electronic health record (EHR) systems to embed the tool within existing clinical workflows.

Additionally, the project has highlighted the critical need to standardize fundus imaging formats and preprocessing protocols to improve cross-dataset compatibility and ensure reliable performance across diverse clinical settings. Further research should also include domain adaptation strategies to enhance generalizability, as well as prospective validation studies to evaluate model performance with real-world patient data prior to deployment in screening programs.

6 Conclusion

Three CNN models were trained on RFMiD, APTOS, and combined datasets for retinal disease classification. Individual dataset models achieved high accuracy (¿93%) and low validation loss (¡0.16), indicating strong potential for clinical deployment in specific contexts. The combined model exhibited lower validation performance, highlighting challenges in multi-dataset integration due to heterogeneity. Developing generalized models requires domain adaptation, data harmonization, and larger, diverse datasets. Future work should explore these approaches to enhance robustness and equity in AI-powered retinal disease screening.

References

Asia Pacific Tele-Ophthalmology Society (APTOS). (2019). *Diabetic retinopathy detection dataset.* https://www.kaggle.com/c/aptos2019-blindness-detection. (Accessed 2025)

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. doi: 10.1001/jama.2016.17216

Lin, C.-Y., Yang, J.-C., Hsieh, M.-J., Lin, C.-H., Shieh, J.-Y. (2020). Multi-task learning for retinal disease classification with improved accuracy. *Computers in Biology and Medicine*, *125*, 103952. doi: 10.1016/j.compbiomed.2020.103952

Rajalakshmi, R., Subashini, R., Anjana, R. M., Mohan, V. (2018). Automated diabetic retinopathy screening and monitoring using digital fundus images. *Journal of Diabetes Science and Technology*, *12*(2), 295–303. doi: 10.1177/1932296817747773

Retinal Fundus Multi-Disease Image Dataset (RFMiD). (2020). *Retinal fundus multi-disease image dataset.* https://www.kaggle.com/datasets/andrewmvd/retinal-fundus-images-for-multi-disease-classification. (Accessed 2025)

Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H. (2017). Applying artificial intelligence to disease staging: Deep learning for diabetic retinopathy. *Ophthalmology Retina*, *1*(4), 322–328. doi: 10.1016/j.oret.2017.03.009