DEEP LEARNING FOR RETINAL DISEASE CLASSIFICATION

COMPARATIVE ANALYSIS OF RFMID, APTOS AND COMBINDED DATASETS

PRESENTED BY: ETANA DISASA
MSDS 692 - DATA SCIENCE PRACTICUM
REGIS UNIVERSITY

INTRODUCTION & PROBLEM STATEMENT

- DIABETIC RETINOPATHY (DR) IS A LEADING CLAUSE OF PREVENTABLE BLINDNESS.
- EARLY DIAGNOSIS IS CRITICAL, BUT OPHTHALMOLOGY RESOURCES ARE LIMITED.
- PROBLEM STATEMENT:
  Can a deep learning model accurately classify retinal fundus images into DR-positive or DR-negative to support early detection?

DATASETS & DATA PREPARATION

- RFMiD: 3,200+ images, already includes binary DR label
- APTOS 2019: 3,662 images, multi-class DR labels (converted to binary)
- Preprocessing:
  - Resized to 224x224
  - Normalized to [0, 1]
  - Augmentations: flip, rotate, shift, zoom
- Combined dataset used for generalization

MODEL ARCHITECTURE & TRAINING

MODEL: MOBILENETV2 WITH CUSTOM TOP LAYERS

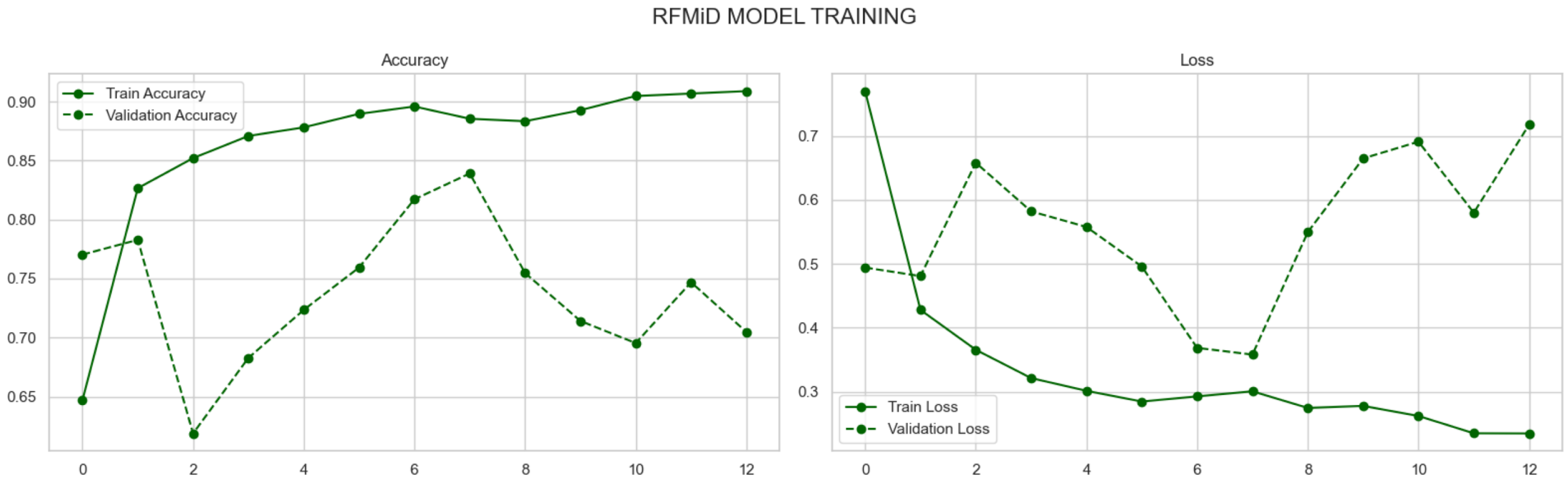
WHY MOBILENETV2?

- Efficient yet accurate for medical image tasks
- Suited for low-resource environments

TRAINING STRATEGIES:

- Dataset-specific models: RFMiD, APTOS, Combined
- Callbacks: EarlyStopping, ReduceLROnPlateau, ModelCheckpoint
- Layer freezing to manage computational load

LEARNING CURVE OF RFMID DATASET



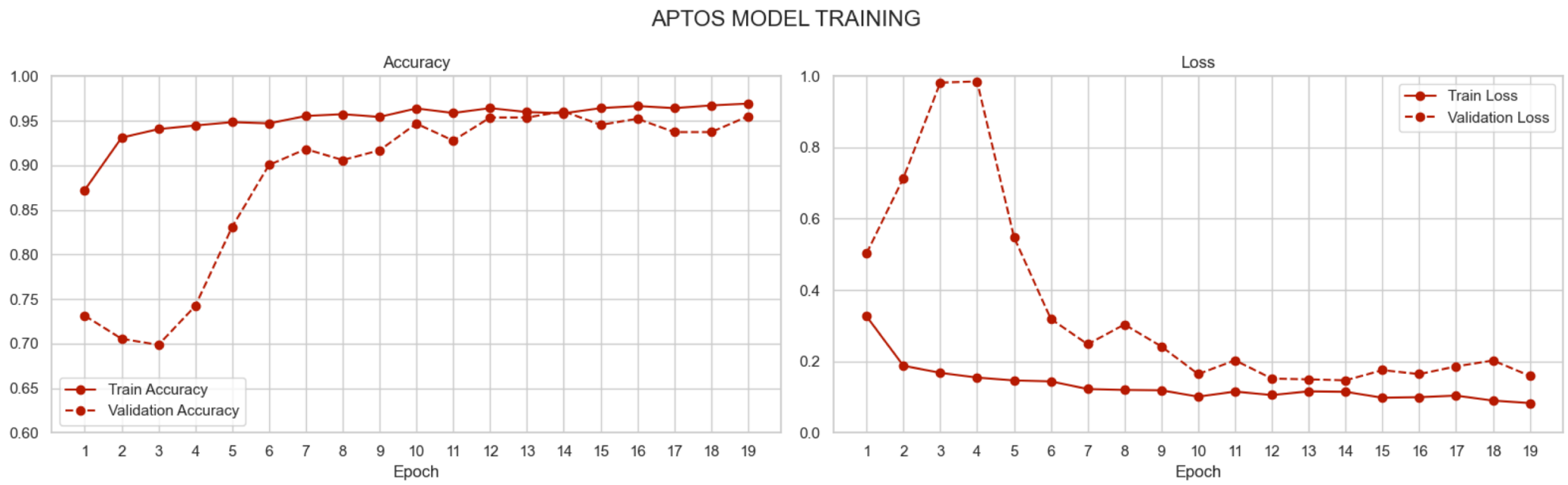
LEARNING CURVE OF RFMID DATASET

- 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.

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

LEARNING CURVE OF APTOS DATASET



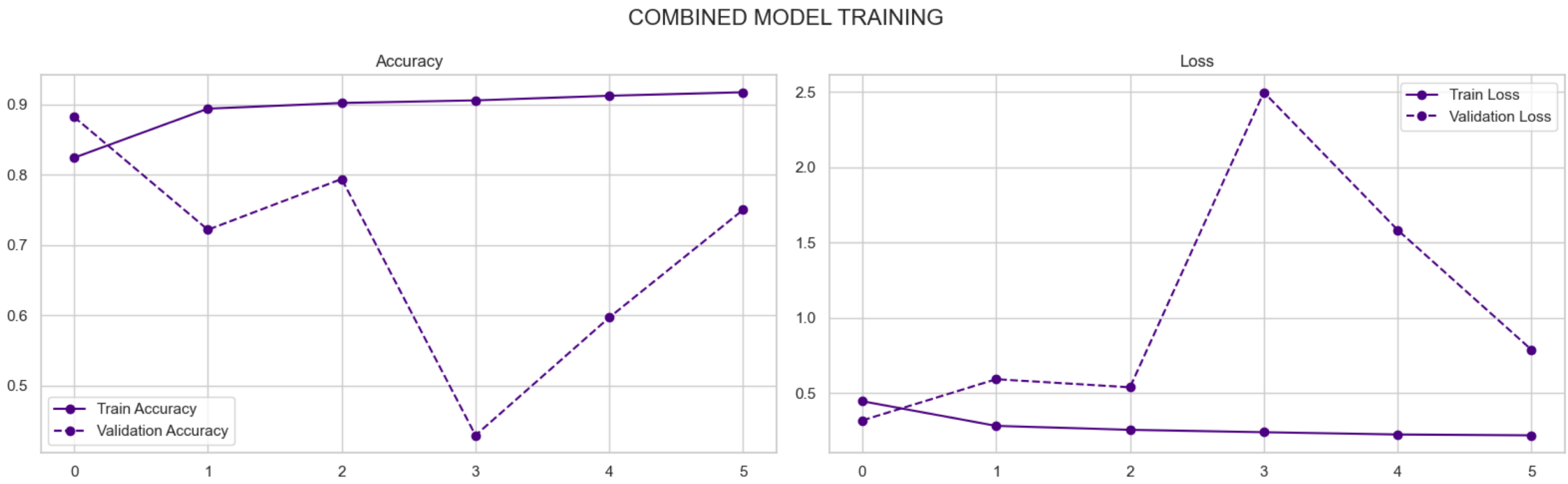
LEARNING CURVE OF APTOS DATASET

- 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.

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	

LEARNING CURVES FOR COMBINED DATASET



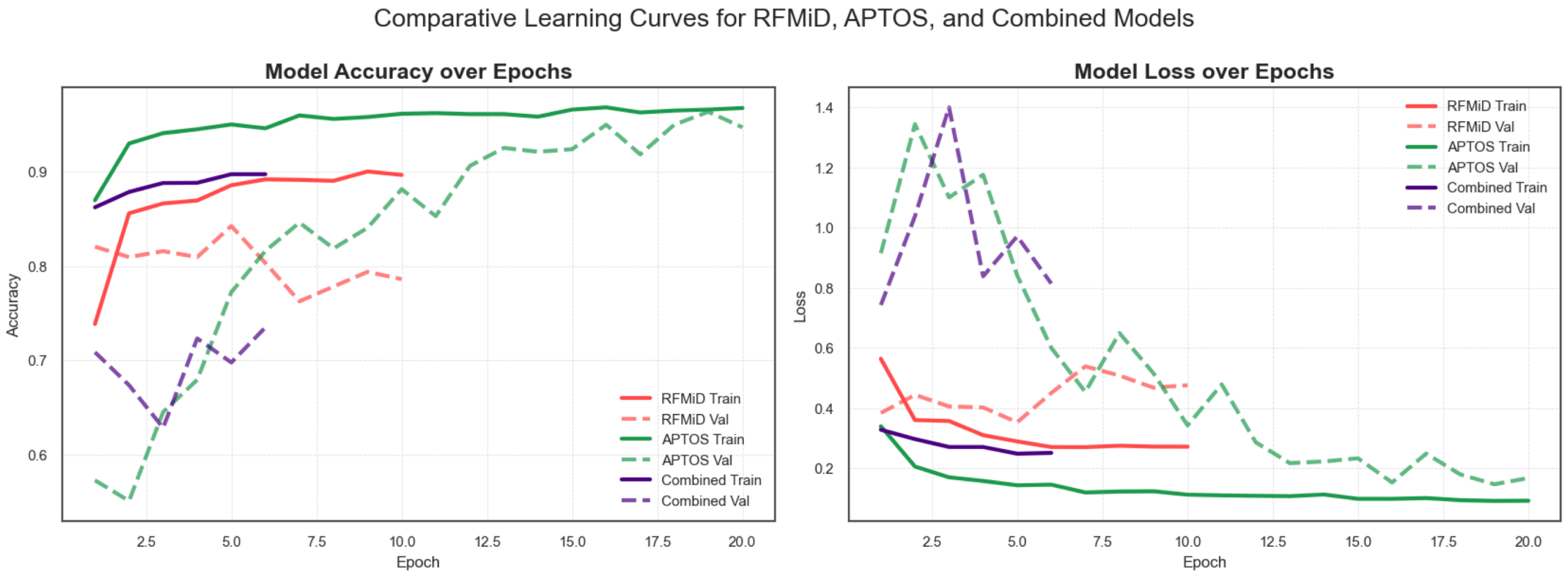
LEARNING CURVES FOR COMBINED DATASET

- 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.

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.

COMPARISON AMONG THE THREE MODELS



COMPARISON AMONG THE THREE MODELS

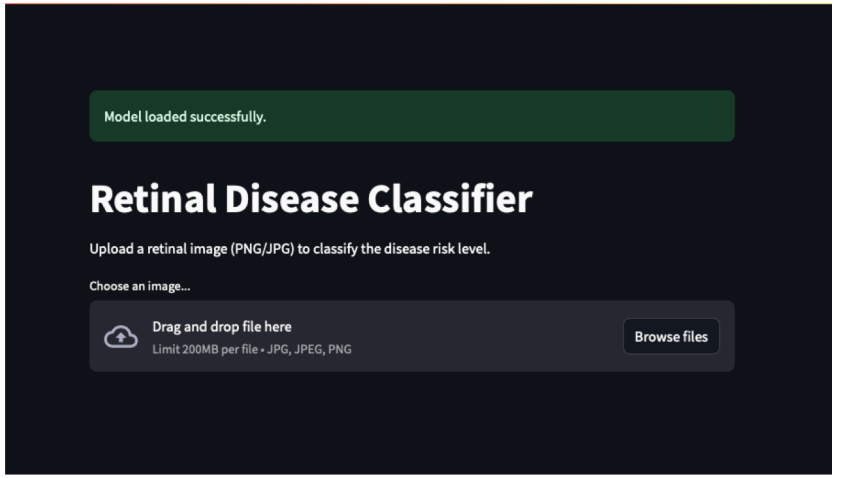
- Models trained on individual datasets outperformed the combined model, suggesting dataset heterogeneity—such as differences in imaging, labels, and patient demographics—impacts performance.
- Naïve combination without domain adaptation or harmonization reduced model generalization and increased overfitting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

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

APPLICATION AND LEARNED LESSONS

- 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.



CHALLENGES ENCOUNTERED

Challenges

- Label mismatch: multi-class vs binary DR labels
- File path inconsistencies between datasets
- Long training times due to limited computing power
- Balancing generalization without sacrificing accuracy

Solutions

- Label harmonization & preprocessing scripts
- Layer freezing + early stopping
- Modular training pipelines

CONCLUSION/FUTURE STEPS

- Combining Datasets Does Not Guarantee Better Performance: This research demonstrated that merging heterogeneous datasets (RFMiD + APTOS) did not improve model accuracy and, in some cases, reduced generalizability due to differences in image quality, label standards, and patient demographics across datasets.
- Lightweight CNNs Enable Effective AI Screening Tools: Using MobileNetV2 with transfer learning achieved high accuracy while remaining computationally efficient, suggesting lightweight architectures are practical for deployment in real-world, resource-constrained clinical environments.
- Bridging Model Development and Deployment Requires Standardization: Developing a Streamlit app prototype confirmed the feasibility of clinical deployment, but limitations in image formats and quality highlighted the need for standardized imaging protocols and robust preprocessing pipelines to ensure reliable AI integration in healthcare workflows.

CONCLUSION/FUTURE STEPS

- Implement Domain Adaptation Techniques: Develop and test domain adaptation or transfer learning methods(e.g. feature alignment, adversarial adaptation) to improve model generalizability when combining heterogeneous retinal datasets.
- Expand to Multi-Class and Multi-Disease Classification: Extend the current binary classification framework to multi-class models capable of detecting and staging multiple retinal diseases, enhancing clinical utility and screening comprehensiveness.
- Integrate Explainable AI (XAI) for Clinical Trust: Incorporate explainability tools such as Grad-CAM heatmaps into the Streamlit app to provide visual explanations of model predictions, increasing clinician trust and facilitating adoption in real-world workflows.

THANK YOU

ETANA DISASA

edisasa@regis.edu