From Data to Diagnostics

Enhancing Early Diabetes Detection & Predicting Diabetic Retinopathy Severity with Machine Learning

---

Ashley Huynh |Hassan Qureshi
Regina Santos | Victoria Tran

Purpose of the Study

Enhance predictive capabilities for early-onset diabetes and by developing advanced prediction models, the research seeks to support early disease detection and track the potential progression of complications like Diabetic Retinopathy

Data Used

- Tabular dataset that offers an overview of various types of Diabetes
- Unstructured dataset comprising of diverse retinal scans that illustrate the severity of Diabetic Retinopathy

Diabetes is a disease that occurs when your blood sugar is too high and your body either doesn't have the ability to produce enough insulin or can't use insulin properly.

Based on International Diabetes Federation 2021 Statisctics ...

01 There are 536.6 million people within the age of 20-79 years old that have diabetes

02 Cases are going to increase up to 642.8 million in 2030 and up to 783.7 million in 2045

03 People affected by this disease risk getting damage to the kidneys, nerves, heart, and eyes

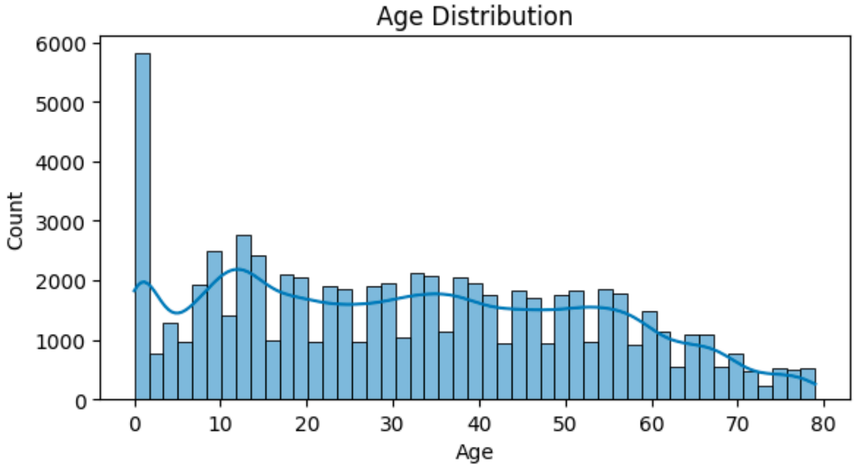
Data Understanding: First Dataset

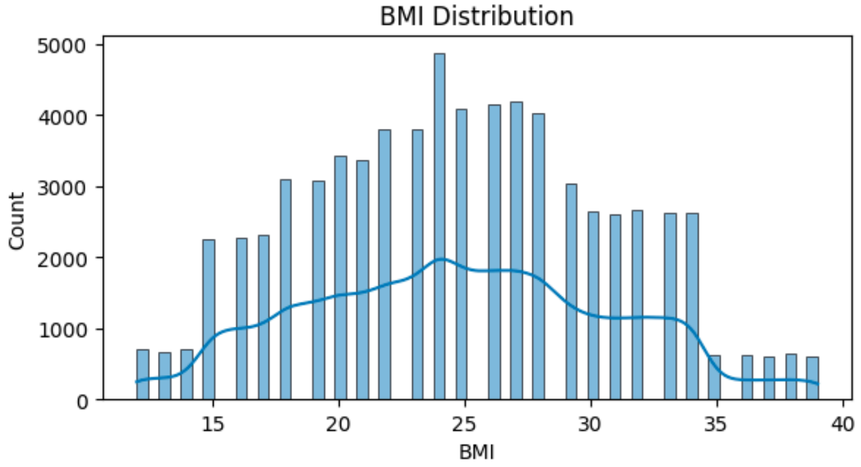
- 70,000 unique records
- 34 features that shows a range of factors influencing diabetes
- Out of the 34, the key features the study focuses on are:
  - Insulin level
  - BMI
  - Blood Pressure
  - Cholesterol level
  - Blood Glucose Level
  - Pancreatic health
- Supervised learning method is performed

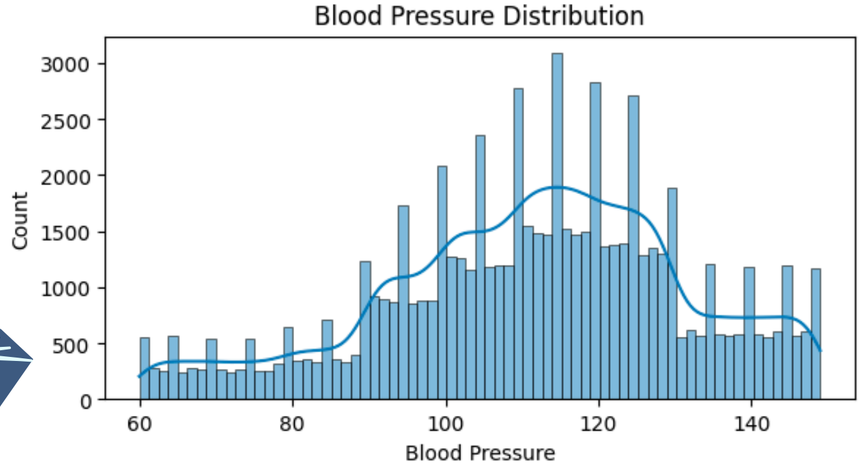
Types of Diabetes

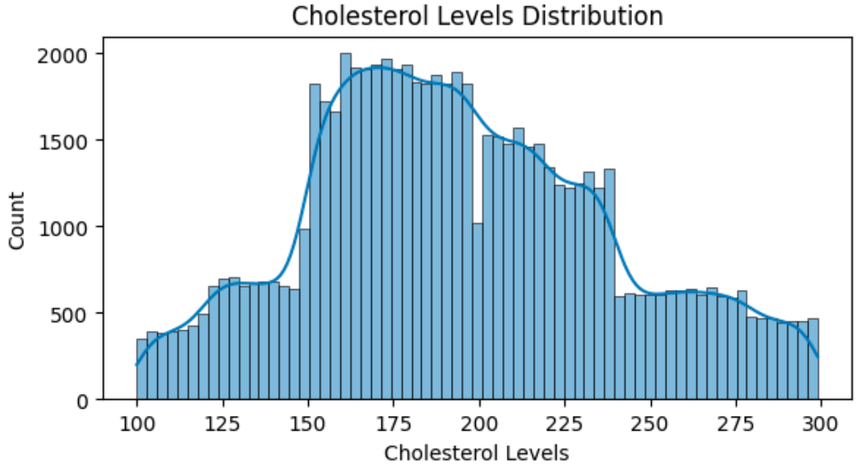
- Maturity-onset Diabetes of the Young (MODY)
- Secondary Diabetes
- Cystic Fibrosis--Related Diabetes (CFRD)
- Neonatal Diabetes Mellitus (NDM)
- Wolcott-Rallison Syndrome
- Wolfram Syndrome
- Type 1 Diabetes
- Type 2 Diabetes
- Type 3C Diabetes (Pancreatic Diabetes)
- Gestational Diabetes
- Steroid-Induced Diabetes
- Latent Autoimmune Diabetes in Adults (LADA)
- Prediabetic

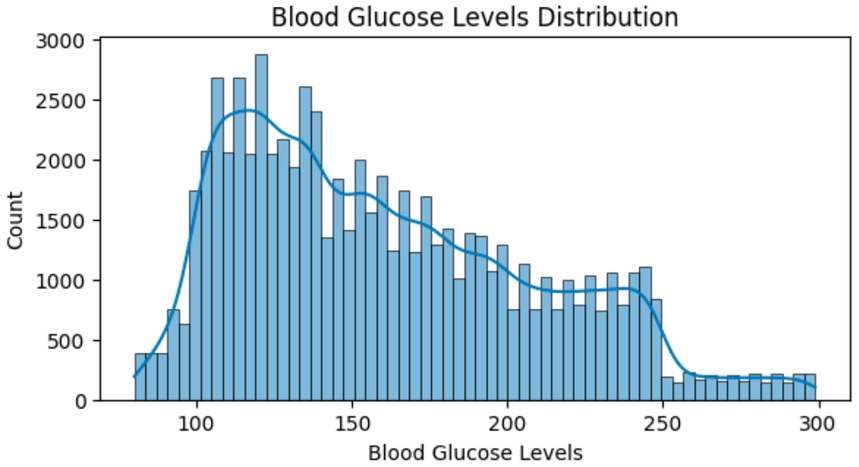
Descriptive Statistics

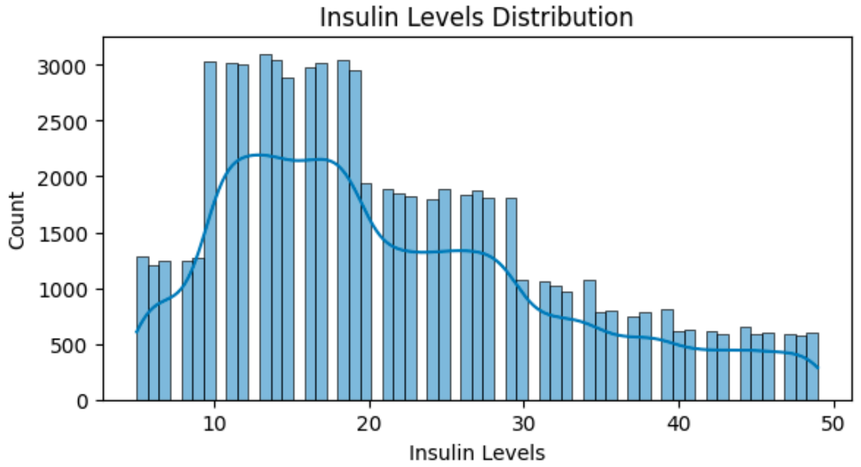










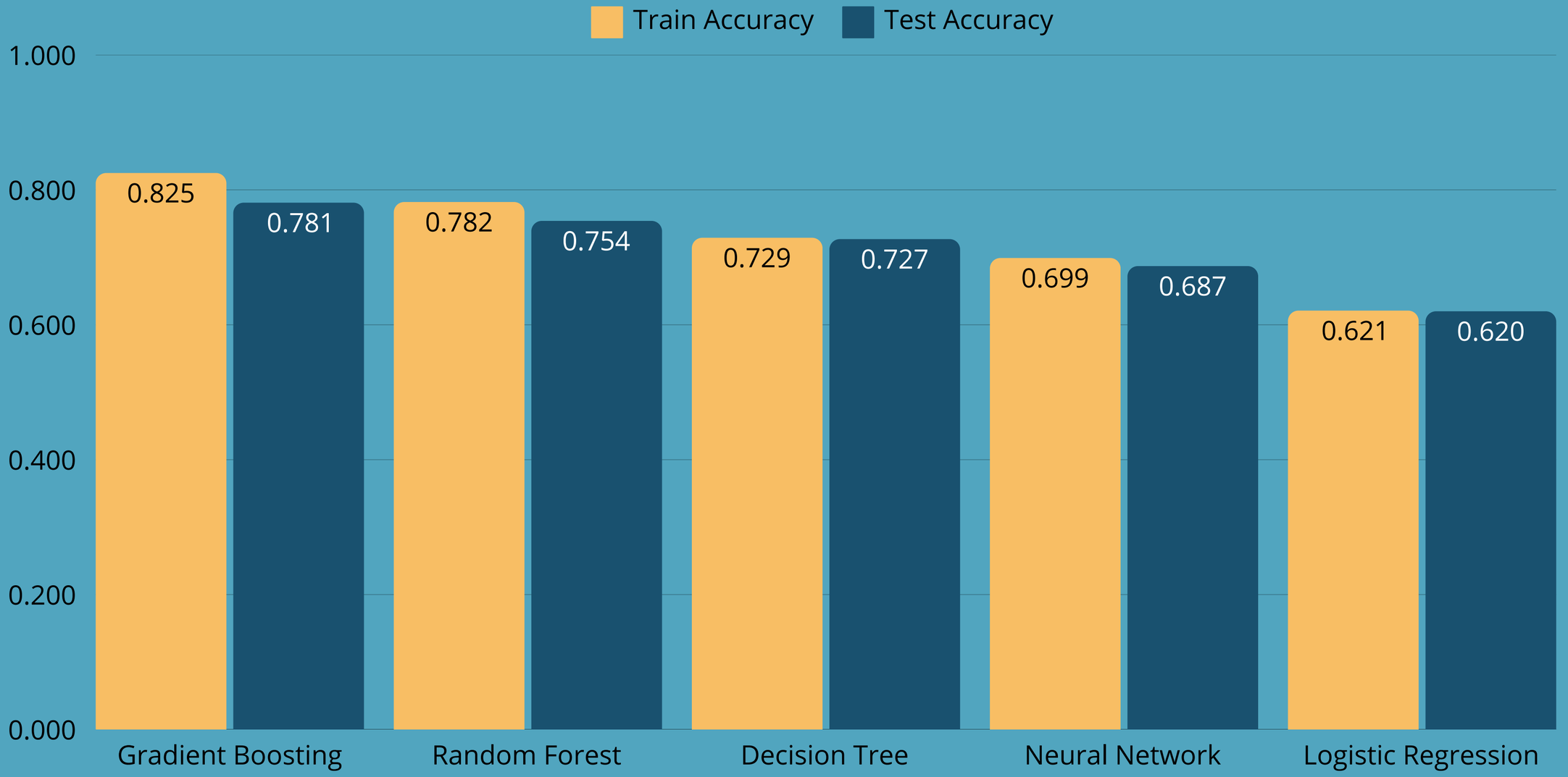


Data Modeling

- Decision Tree Classifier
- Logistic Regression
- Random Forest
- Gradient Boosting Classifier
- Neural Network

Model Comparison

Machine Learning Model Comparison by Overall Accuracy



Best Model Confusion Matrix

% of classification are correctly labeled

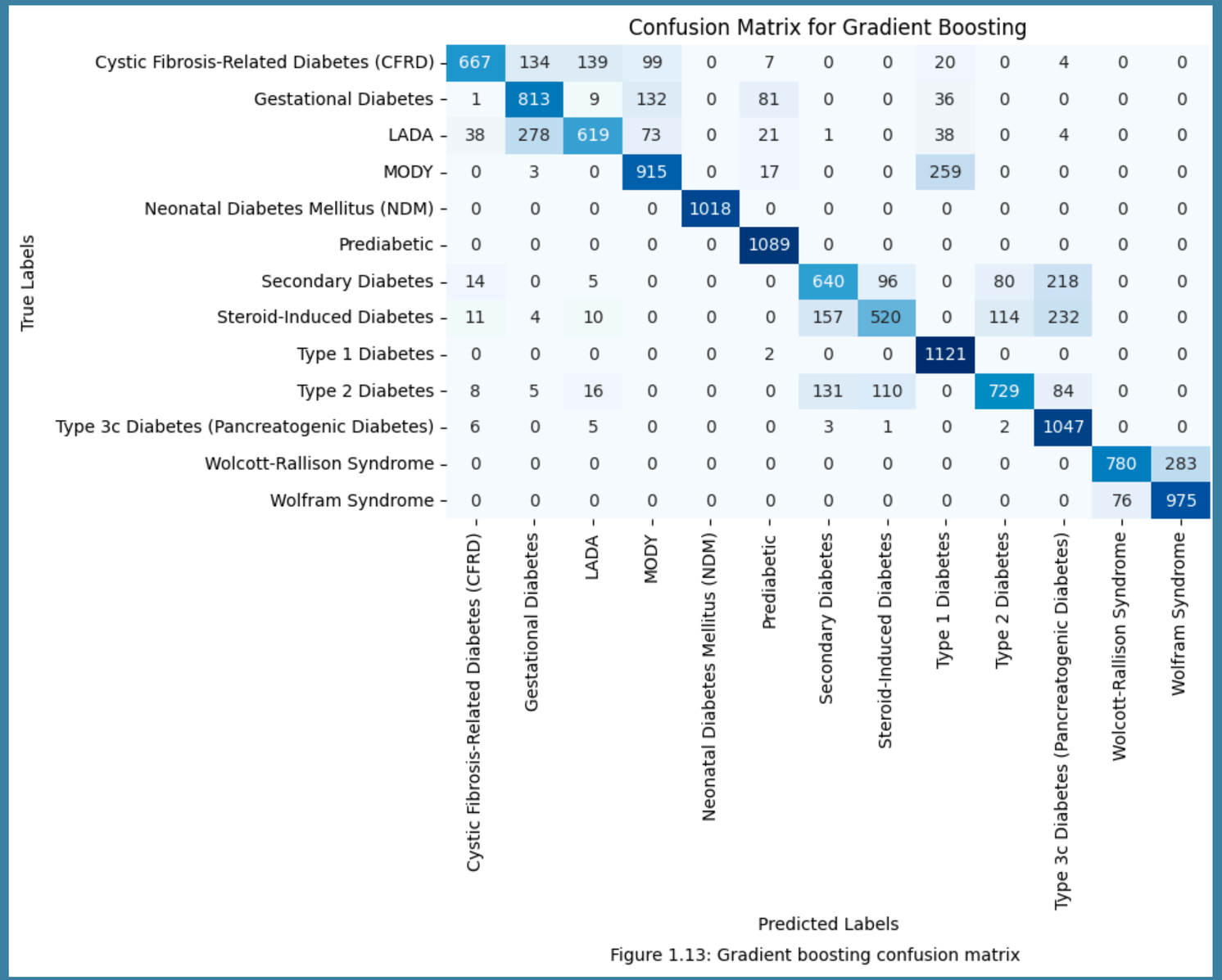
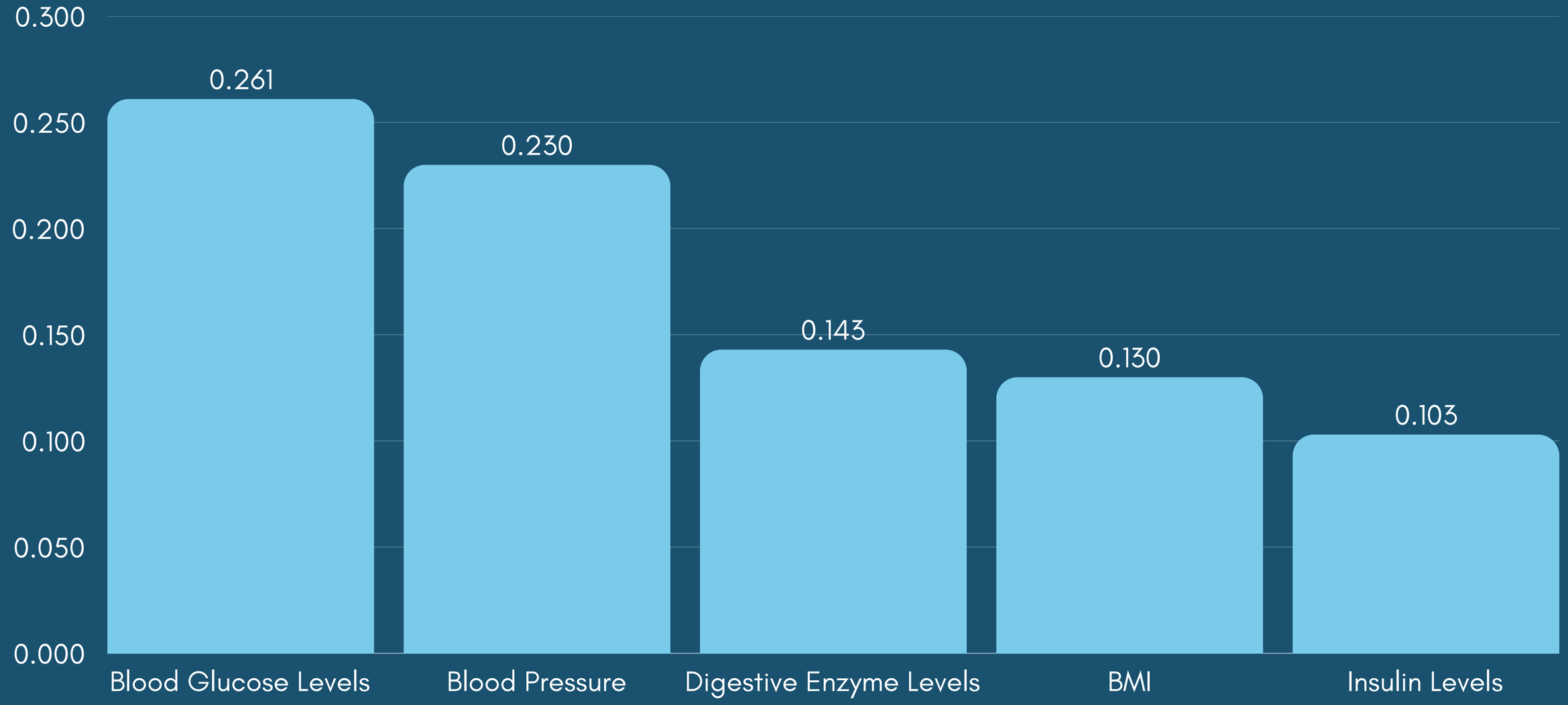


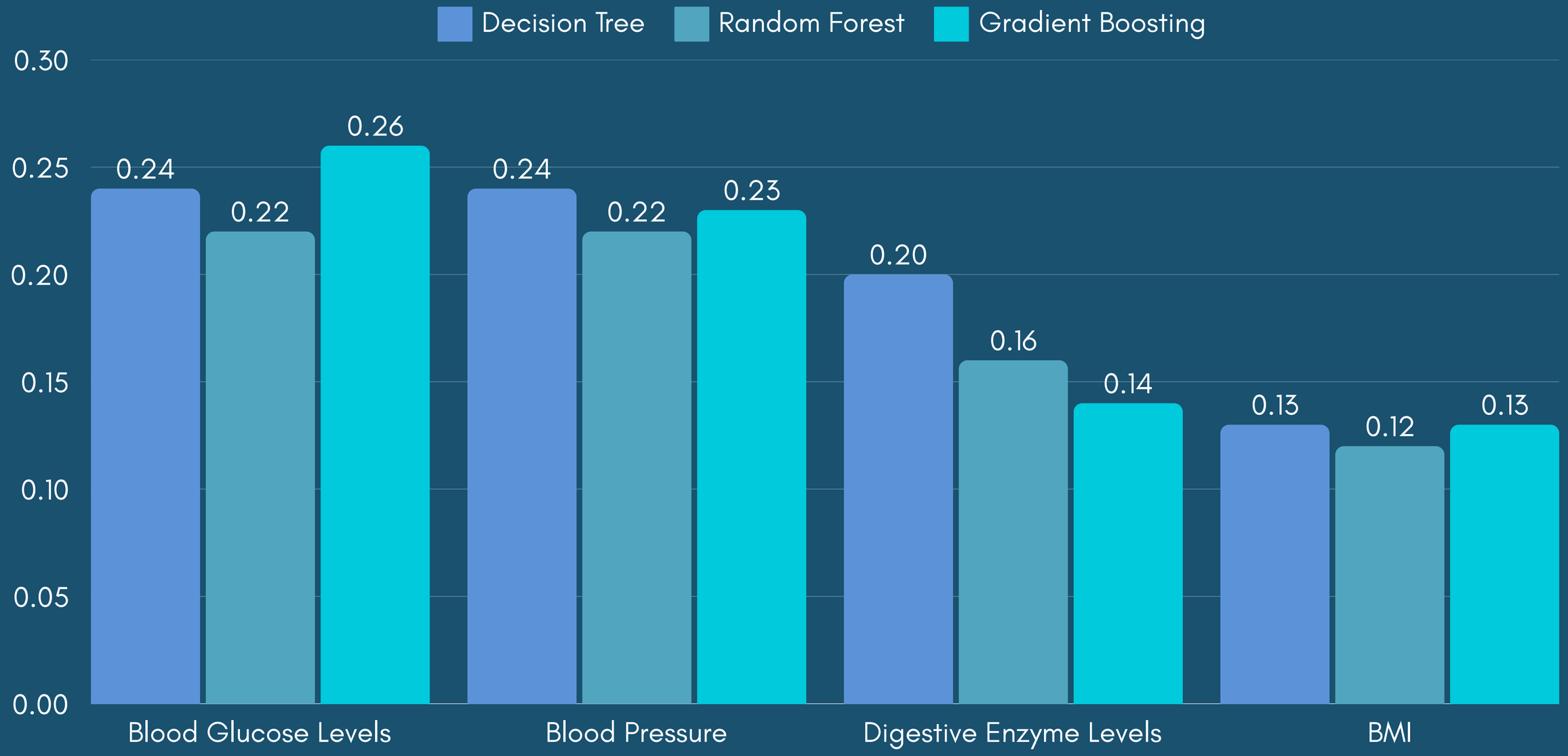
Figure 1.13: Gradient boosting confusion matrix

Top 5 Features Importance

*Gradient Boosting



Top 4 Overall Features Importance

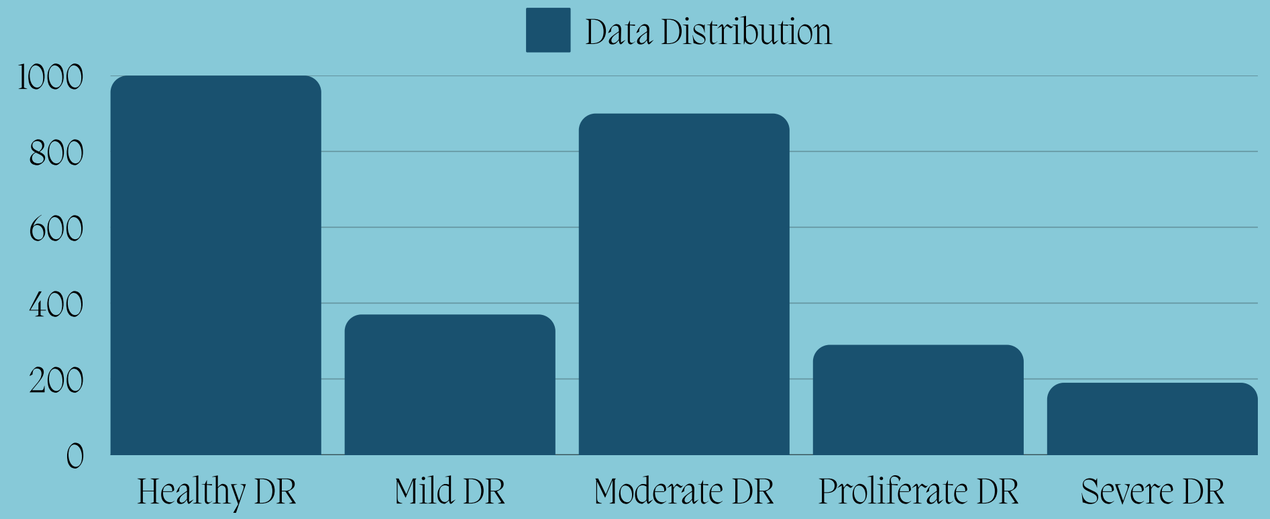


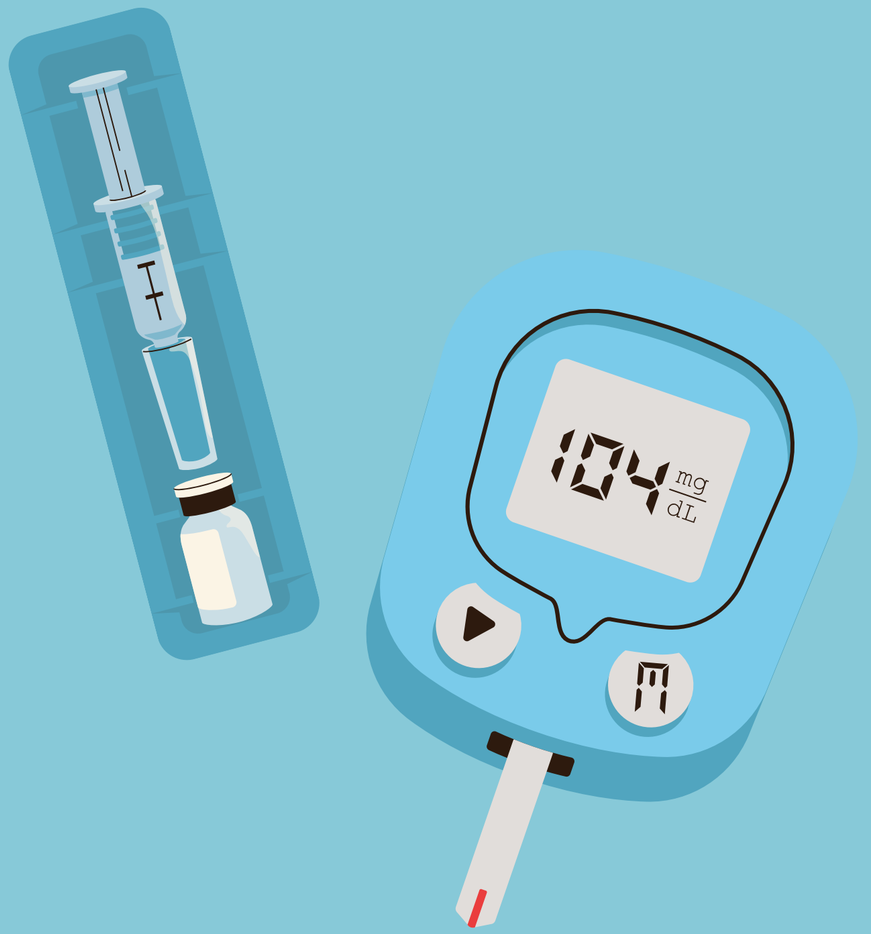
“

Diabetic Retinopathy (DR) is a diabetes-related eye condition that affects the blood vessels in the retina

Data Overview:

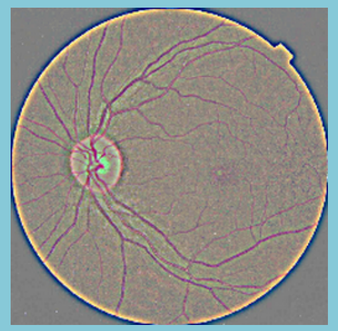
- This dataset contains a total of 2,750 retinal scan images
- Diabetic Retinopathy Classes:
  - Healthy cases
  - Mild cases
  - Moderate cases
  - Proliferative cases
  - Severe cases



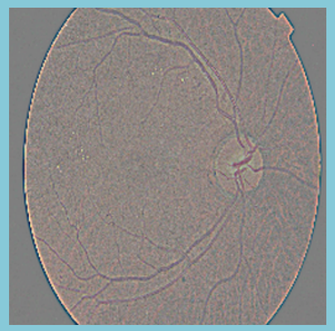


Diabetic Retinopathy Classification

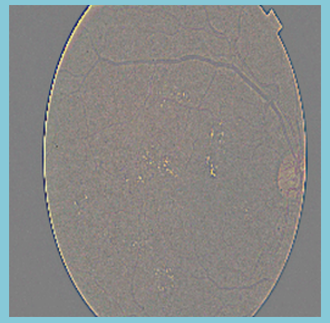
Healthy
No abnormalities present



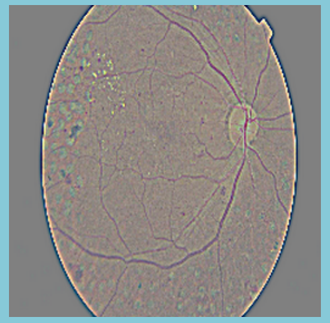
Mild
Some microaneurysms can be seen



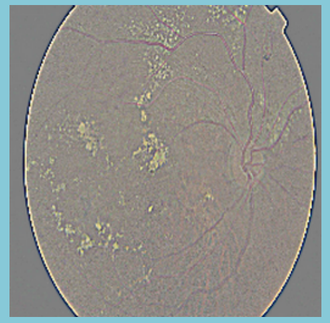
Moderate
Microaneurysms are more prominent



Proliferate
Neovascularization and preretinal hemorrhage can be observed



Severe
Shows intraretinal hemorrhages, venous beading, and prominent intraretinal abnormalities



Data Pre-Processing

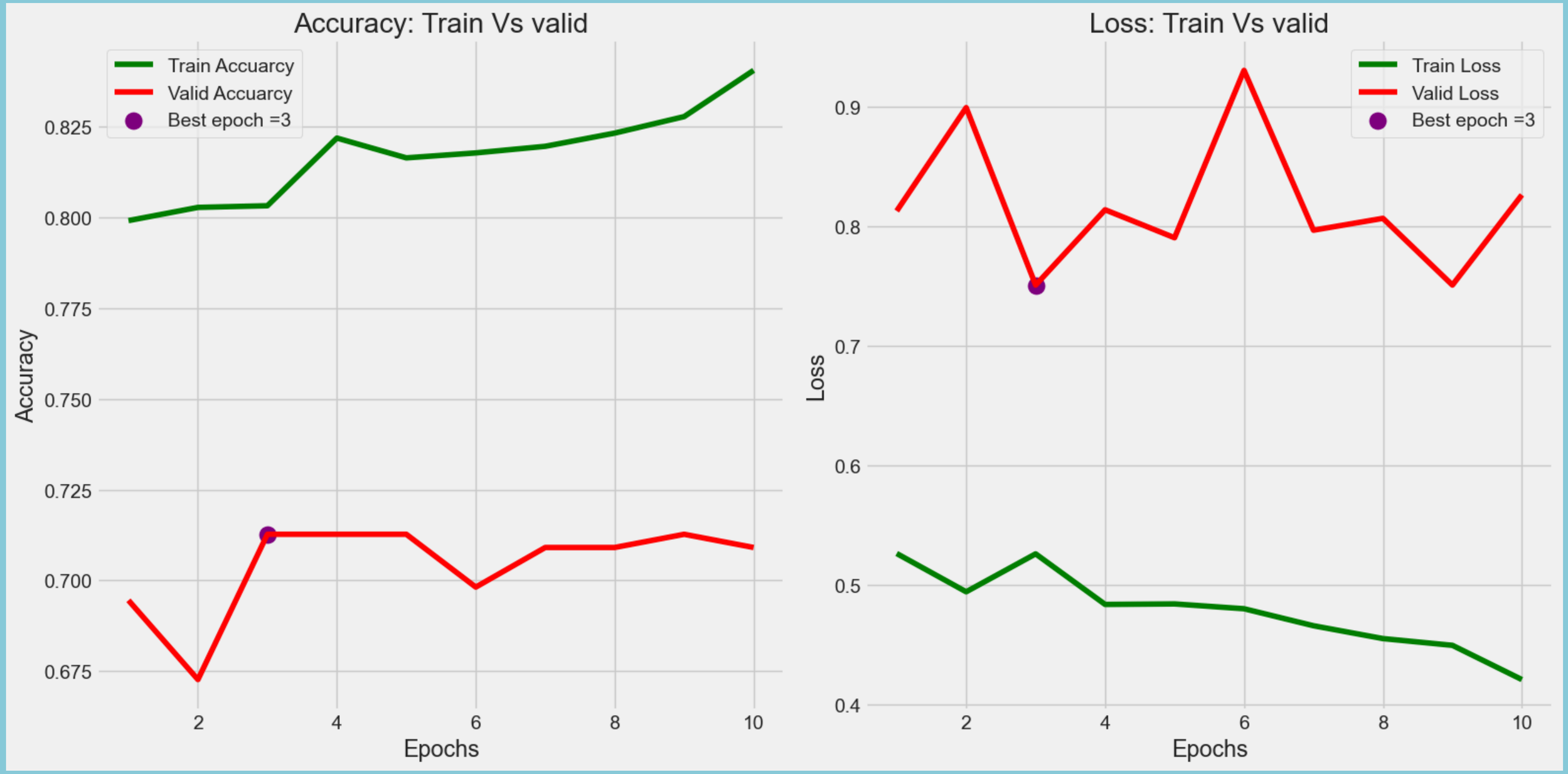
- Image Paths and Labels Extraction
- Data Conversion
- Dataset Splitting
  - Training Set: 80% of the data
  - Test Set: 10% of the data
  - Validation Set: 10% of the data
- Data Augmentation
  - Random rotations, nearest fill
- Resizing
- Batch Processing
- Visualization

Model Implementation

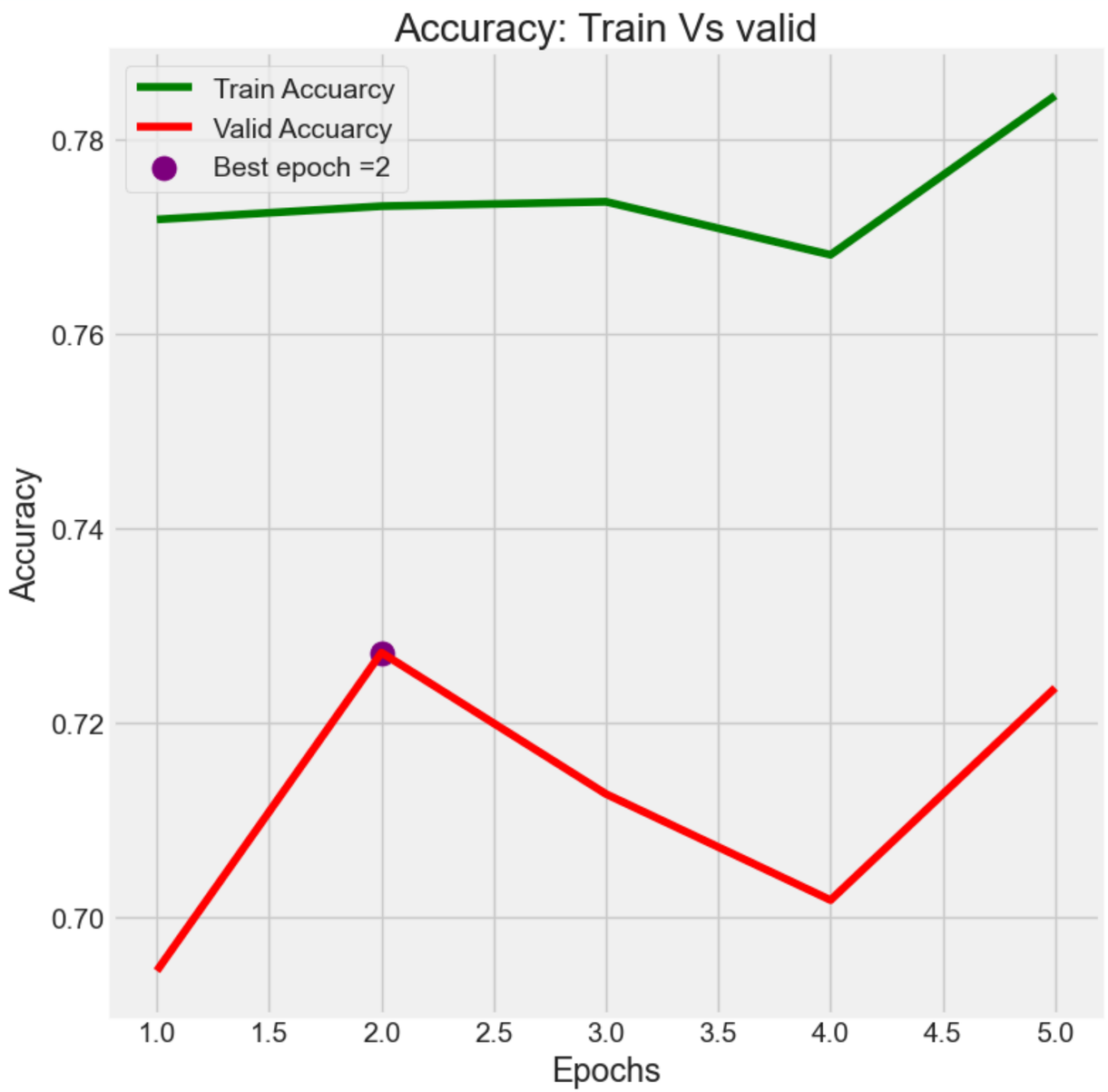
- Neural Network: CNN
- Model: EfficientNB3
- Activation: ELU/Softmax
- Optimizer: AdaMax (lr: 0.0001)
- Loss: Categorical Cross Entropy

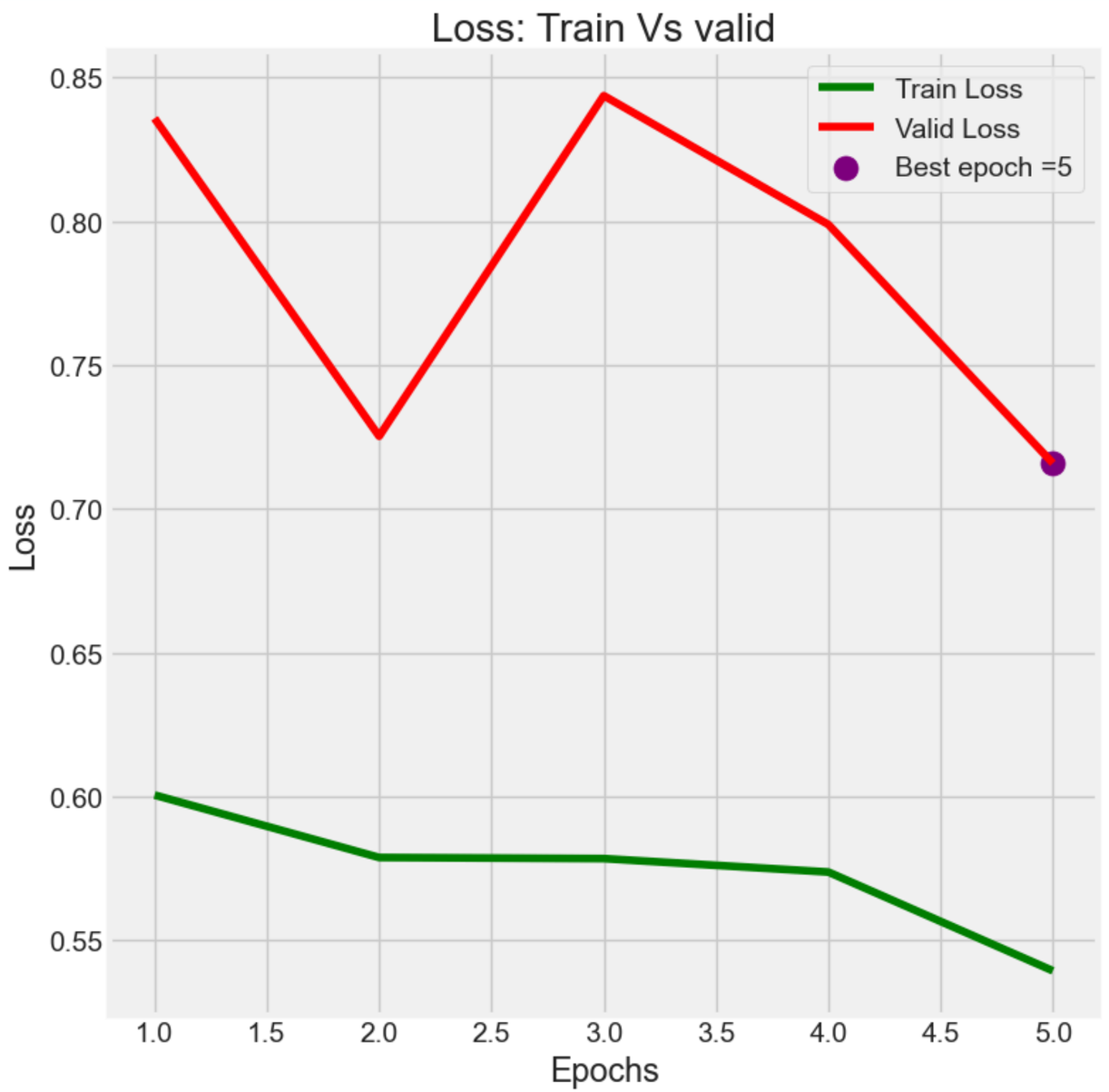
Layer	Output Shape	Purpose
Input	(224, 224, 3)	Original RGB input image
EfficientNetB3	(None, 7, 7, 1536)	Extracts high-level visual features
Dropout	(None, 7, 7, 1536)	Prevents overfitting during training
Flatten	(None, 75264)	Reshapes into a 1D vector for dense layers
Dense (elu)	(None, 512)	Fully connected layer with 512 neurons
Dense (elu)	(None, 256)	Reduces features to 256 units
Dense (elu)	(None, 128)	Further compressing into 128 units
Output (Softmax)	(None, 5)	Class probabilities for diabetic retinopathy severity levels

Training and Valid Scores: 10 Epochs

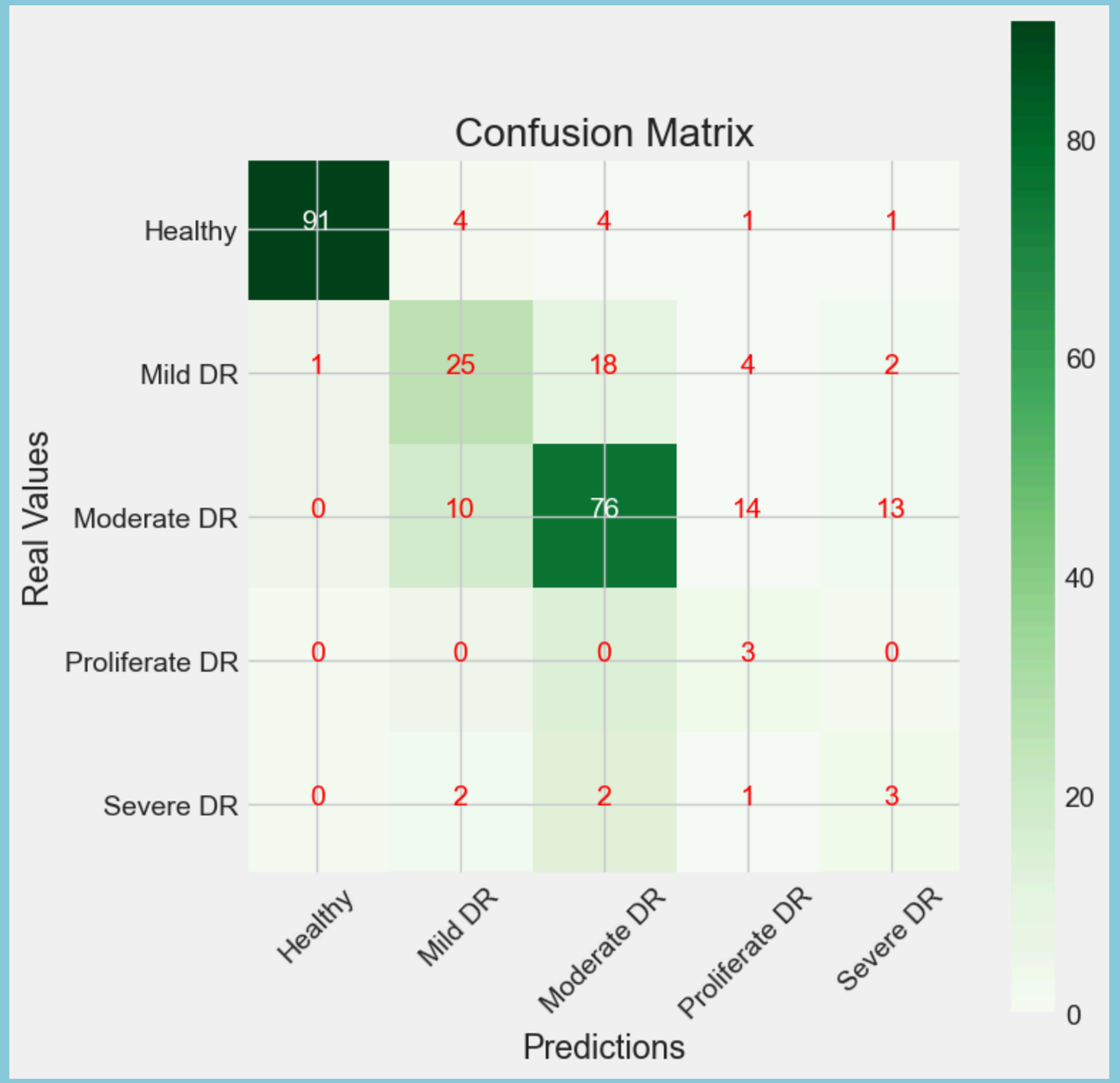


Training and Valid Scores: 5 Epochs

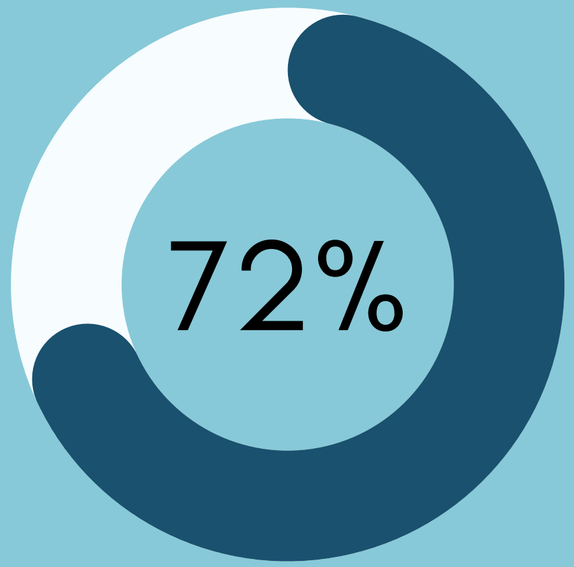




Prediction Confusion Matrix



Classification Report



Overall Accuracy

	precision	recall	f1-score	support
Healthy	0.90	0.99	0.94	92
Mild DR	0.50	0.61	0.55	41
Moderate DR	0.67	0.76	0.71	100
Proliferate DR	1.00	0.13	0.23	23
Severe DR	0.38	0.16	0.22	19
accuracy			0.72	275
macro avg	0.69	0.53	0.53	275
weighted avg	0.73	0.72	0.69	275

Results

Dataset 1

Descriptive Statistics

Diabetes can happen to anyone at any age. There are numerous factors that affect diabetes type and level of its severity.

Best Performance: Gradient Boosting

Machine learning model that use ensemble learning to predict diabetic results by focusing on minimizing errors of the previous model in the sequence. **Accuracy = 78%**

Feature Importance: Blood Glucose Levels

Most correlated variable predicting diabetic result. Managing blood glucose level since high blood glucose level damage the blood vessels in your eyes overtime.

Dataset 2

EfficientNetB3 for DR Classification

We used the EfficientNetB3 model to classify the severity of DR because of its compound scaling properties and transfer learning benefits through pre-trained pixel feature weights.

Dataset Bias and Preprocessing Choices

We decided not to apply over/under-sampling techniques to our dataset, thus, the model showed a bias towards healthy images, which is typical in real-world data

Training Results and Practical Impact

After training the model for multiple epochs, we observed accuracy stabilizing around **~72%**, which is fairly good as a means to assist ENT physicians in identifying DR severity in diabetic patients

In Conclusion ...

- **Model Utilization**
  - Integrating the model into a clinical decision support system to provide insights to physicians
  - Researchers could leverage the model to gain deeper insights into the relationship between diabetes and vision health
- **Future Improvements**
  - Hyperparameter tuning can improve the accuracy of the models
  - Continuously monitor and update the model as new data becomes available

Thank you!!