**Glaucoma Detection from Retinal Fundus Images**

**ML Project**

**Team Members:**

1. Mohamed Habiba

2- Ahmed Tarek Aboughanima

3- Mohamed Atef Omara

4- Mohamed Abdelsalam Farag Gwida

5- Menna Hesham Eraky

**1. Executive Summary**

**Objective:** Develop and compare multiple model architectures for accurate binary classification of retinal fundus images (healthy vs. glaucomatous), facilitating early detection and screening in clinical and resource-constrained environments.

**Dataset:** The study utilized a consolidated dataset of **8,621 retinal fundus images** (comprising 5,293 healthy and 3,328 glaucomatous cases), aggregated from publicly available sources including the Kaggle fundus collection, ORIGA, and REFUGE. The data was stratified into training (70%), validation (15%), and testing (15%) sets to ensure robust evaluation.

**Methodology:** The project conducted a dual-pipeline investigation:

1. **Classical Machine Learning:** A "white-box" approach evaluating Logistic Regression, KNN, SVM, Random Forest, and XGBoost. This phase emphasized explicit feature engineering, extracting domain-specific descriptors (LBP, HOG, Cup-to-Disc Ratio) to benchmark interpretability and performance on limited hardware.
2. **Deep Learning (Transfer Learning):** A "black-box" approach leveraging state-of-the-art Convolutional Neural Networks (CNNs) pre-trained on ImageNet. Architectures including **VGG16**, **ResNet50**, and **DenseNet121** were fine-tuned to automatically learn hierarchical feature representations from raw fundus images.

**Key Results:** The investigation confirmed that Deep Learning architectures provide superior diagnostic precision compared to classical methods. The **DenseNet121** model emerged as the optimal solution, effectively handling the subtle textural and structural patterns associated with glaucoma.

* **Best performing model:** **DenseNet121 (Transfer Learning)**
* **Test Accuracy:** **95.4%** *(vs. 88.0% for the best Classical ML model)*
* **Sensitivity (Recall):** **97.2%** *(Critical for minimizing false negatives in medical screening)*
* **AUC-ROC:** **0.981** *(Indicates exceptional class separation capability)*
* **Outcome:** Successfully developed a high-performance diagnostic tool capable of automated screening, alongside clinically interpretable visualization tools (Grad-CAM/Feature Importance plots) to support medical decision-making.

**2. Project Overview**

**2.1 Introduction**

Glaucoma represents a constellation of optic neuropathies characterized by progressive optic nerve damage, resulting in irreversible vision loss. As the second leading cause of global blindness, glaucoma affects over 80 million individuals worldwide. Its particularly insidious characteristic—asymptomatic progression until substantial permanent damage occurs—renders early detection critically important yet challenging.

**2.2 Project Objectives**

**Primary Objective**

Develop and systematically compare state-of-the-art model architectures for accurate binary classification of retinal fundus images into healthy and glaucomatous categories.

**Secondary Objectives**

1. **Model Performance Optimization**
   * Maximize sensitivity (recall) to minimize false negatives—clinically, undetected glaucoma cases pose greater risk than false positives
   * Maintain appropriate specificity to control false positive rates
   * Optimize F1-Score for balanced precision-recall performance
2. **Comprehensive Architecture Comparison**
   * Evaluate traditional machine learning models and CNN architectures
   * Compare transfer learning implementations (ResNet50, EfficientNet, VGG16, DenseNet, Xception)
   * Analyze performance trade-offs considering accuracy, computational efficiency, and interpretability

**2.3 Expected Outcomes**

**Technical Deliverables:**

* Robust, validated glaucoma classification system
* Comparative analysis identifying optimal architecture for this diagnostic task
* Explainability frameworks that enhance clinician confidence and enable verification

**Research Contributions:**

* Benchmark comparison of diverse model architectures for glaucoma detection
* Insights regarding transfer learning efficacy in medical imaging applications
* Open-source implementation for reproducibility and future research extension

**2.4 Project Scope**

*Models:*

* Multiple classification machine learning algorithms
* Custom CNN architectures
* Transfer learning leveraging ImageNet pre-trained models

*Analysis:*

* Comprehensive performance metrics (accuracy, precision, recall, F1-score, AUC-ROC)
* Confusion matrix interpretation
* Error pattern characterization
* Feature visualization techniques

**3. Theoretical Foundations**

**3.1 Logistic Regression**

Logistic Regression serves as a foundational linear model for binary classification tasks, such as distinguishing between healthy and glaucomatous fundus images. Despite its simplicity, it provides a strong probabilistic baseline and is highly interpretable.

**Theoretical Foundation and Mathematical Formulation:**

The core idea is to model the probability that a given input image (represented by its feature vector x) belongs to the positive class (Glaucoma, y=1). It applies a linear combination of the input features and then squashes the output through the Sigmoid activation function to map it to a probability between 0 and 1.

* Linear Component: z = w.x + b Where w is the weight vector, x is the feature vector (e.g., flattened image pixels or extracted features), and b is the bias term.
* Sigmoid Activation: P(y=1 | x) = 1 / (1 + e^-z) This S-shaped function converts the linear output z into a probability.

**Model Training and Loss Function:**

The model is trained by optimizing the weights (w) to maximize the likelihood of the observed data. This is equivalent to minimizing the Log Loss (Binary Cross-Entropy):

J(w) = -(1/N) \* Sum [ y \* log(p) + (1-y) \* log(1-p) ]

Where p is the predicted probability for the sample, and y is the true label.

**Regularization for Glaucoma Detection:**

To prevent overfitting—especially critical with high-dimensional image data—regularization techniques are employed:

* L1 Regularization (Lasso): Adds a penalty equal to the absolute value of the weights. It can drive some feature weights to exactly zero, performing automatic feature selection.
* L2 Regularization (Ridge): Adds a penalty equal to the square of the weights. It shrinks all weights proportionally, leading to a model with typically smaller, more distributed weights.

Application to Fundus Images (Experiment 1): In this project, Logistic Regression was utilized primarily in Experiment 1 (Baseline). The raw fundus images were resized and their pixel intensities were flattened into a 1D feature vector.

* **Role**: It served as a benchmark to assess the linear separability of the data.
* **Outcome**: Its performance highlighted that raw pixel data contains complex, non-linear patterns (such as subtle optic cup dilation) that a linear decision boundary cannot capture effectively without advanced feature engineering.

**3.2 K-Nearest Neighbors (KNN)**

K-Nearest Neighbors is a non-parametric, instance-based learning algorithm used for classification. It operates on a simple principle: an unlabeled fundus image is classified by a majority vote of its k most similar (nearest) labeled images in the feature space.

**Theoretical Foundation:**

KNN makes no explicit assumptions about the underlying data distribution. Instead, it "memorizes" the entire training dataset. The prediction for a new sample is derived directly from the training instances closest to it.

**Key Components and Mathematics:**

1. Distance Metric: The definition of "nearest" is critical. Common metrics include:
   * Euclidean Distance: d(x, y) = Square Root of Sum((xi - yi)^2) This is the standard metric used for continuous features in our project.
   * Manhattan Distance: d(x, y) = Sum(|xi - yi|) The choice of metric significantly impacts the model's geometry and performance.
2. Hyperparameter k: The number of neighbors to consider.
   * A small k (e.g., k=1) leads to a complex, noisy decision boundary prone to overfitting.
   * A large k leads to a smoother boundary but may oversimplify and underfit, potentially missing local patterns crucial for glaucoma detection.
3. Voting Scheme: For classification, the class label is assigned by majority vote among the k neighbors. Weighted voting (where closer neighbors have a larger influence) can also be applied.

Application to Fundus Images (Experiment 2): KNN's performance is heavily dependent on the feature representation.

* Challenges: In the baseline experiment (raw pixels), KNN suffered from the "curse of dimensionality," where distance measures become less meaningful in high-dimensional pixel space.
* Optimization: In Experiment 2, we improved KNN's efficacy by feeding it Engineered Features (Texture via LBP and Shape via HOG). By reducing the image to a compact set of meaningful descriptors, the Euclidean distance metric became effective at grouping similar glaucomatous patterns together.

**3.3 Support Vector Machine (SVM)**

Support Vector Machine is a powerful discriminative classifier that finds the optimal hyperplane to separate data from different classes with the maximum possible margin, making it robust for high-dimensional spaces like image feature vectors.

**Theoretical Foundation and Linear SVM:**

For a linearly separable dataset, the goal is to find the separating hyperplane (w.x + b = 0) that maximizes the margin—the distance between the hyperplane and the nearest data points from either class, called support vectors.

* Decision Function: f(x) = sign(w.x + b)
* Optimization Objective: Minimize the squared norm of weights (||w||^2) subject to correct classification constraints.

**Handling Non-Linearity (The Kernel Trick):**

Fundus image data is rarely linearly separable. SVM handles this using the kernel trick, which implicitly maps the input features into a higher-dimensional space where a linear separation is possible.

* Radial Basis Function (RBF) Kernel: K(x, y) = exp(-gamma \* ||x - y||^2) This is the most common kernel for image data, as it can model complex, non-linear relationships. The parameter gamma controls the influence of a single training example.

**Soft-Margin SVM for Overlapping Classes:**

Real-world medical data, like glaucoma images, often has overlapping features. Soft-margin SVM introduces a slack variable and a regularization parameter C to the objective.

* Hyperparameter C: Controls the trade-off between maximizing the margin and minimizing classification error. A large C penalizes misclassifications heavily, leading to a harder margin. A smaller C allows more margin violations, promoting a softer, potentially more generalizable boundary.

Application to Fundus Images (Experiment 3): SVMs are effective when paired with informative feature descriptors.

* ROI Detection: In Experiment 3, we applied SVM to HOG features extracted from the Optic Disc ROI.
* Synergy: The combination of HOG (which captures edge directions) and SVM (which maximizes margins between classes) proved highly effective for distinguishing the vertical elongation of the optic cup, a key indicator of Glaucoma. The RBF kernel allowed the model to handle the non-linear variations in cup shape across patients.

**3.4 Random Forest (RF)**Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (for classification) of the individual trees. It is renowned for its robustness, accuracy, and resistance to overfitting.

**Theoretical Foundation: Ensemble of Decision Trees**

A single decision tree partitions the feature space (e.g., pixel or texture features) using a series of hierarchical, axis-aligned splits based on feature values. While interpretable, single trees are highly prone to overfitting.

**Random Forest introduces two key sources of randomness to build diverse, de-correlated trees:**

1. Bootstrap Aggregating (Bagging): Each tree is trained on a random subset (with replacement) of the training data. This creates variation in the data each tree sees.
2. Random Feature Selection: At each split in a tree, the algorithm considers only a random subset of all available features. This prevents strong, correlated features from dominating all trees.

**Prediction and Key Characteristics:**

* Aggregation: For a new fundus image, each tree in the forest "votes" for a class (Healthy or Glaucoma). The final prediction is the majority vote.
* Inherent Advantages:
  + Reduced Overfitting: The ensemble average of many noisy but unbiased trees results in a model with lower variance than a single tree.
  + Feature Importance: Provides a native measure of which features were most influential in making predictions, offering valuable clinical interpretability.

**Hyperparameters for Optimization:**

* n\_estimators: Number of trees in the forest.
* max\_depth: The maximum depth of each tree (controlling model complexity).
* class\_weight: Can be set to 'balanced' to automatically adjust weights inversely proportional to class frequencies.

Application to Fundus Images (Experiment 4): Random Forest excelled across our pipelines due to its ability to handle mixed feature types.

* Handling Imbalance: We utilized class\_weight='balanced' to address the dataset imbalance (5,293 Healthy vs 3,328 Glaucoma).
* Feature Selection: In Experiment 4, Random Forest allowed us to integrate various physiological features. Its Feature Importance attribute was crucial for analyzing which specific texture (LBP) or shape (HOG) features contributed most to the diagnosis, adding a layer of explainability to the black-box classification.

**3.5 XGBoost (Extreme Gradient Boosting)**

XGBoost is an optimized implementation of gradient boosting machines, designed for high efficiency and scalability in classification tasks, such as detecting glaucoma from fundus images. It builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessors, leading to a powerful model that often outperforms other ensemble methods like Random Forest in structured data scenarios.

**Theoretical Foundation: Gradient Boosting Framework**

XGBoost is rooted in the gradient boosting algorithm, which constructs a strong predictive model by combining multiple weak learners (typically shallow decision trees) in an additive manner. Unlike bagging-based ensembles (e.g., Random Forest), boosting focuses on iteratively improving the model by emphasizing misclassified instances from previous iterations.

The core concept is to approximate the target function (e.g., the probability of glaucoma) as a sum of base models: y\_hat\_i = sum\_{k=1 to K} f\_k(x\_i)

Where y\_hat\_i is the predicted label for input x\_i, K is the number of trees, and each f\_k is a decision tree added at iteration k.

**Mathematical Formulation and Optimization**

XGBoost minimizes a regularized objective function that balances predictive accuracy and model complexity. For binary classification (e.g., Glaucoma vs. Normal), it uses logistic loss as the base, combined with regularization terms: L = sum\_{i=1 to n} l(y\_i, y\_hat\_i) + sum\_{k=1 to K} Omega(f\_k)

* Loss Function l: Binary logistic loss, l(y, y\_hat) = - [y log(p) + (1-y) log(1-p)], where p = sigma(y\_hat) is the sigmoid-transformed prediction.
* Regularization Omega(f): Omega(f) = gamma T + (1/2) lambda ||omega||^2 + alpha ||omega||\_1, where T is the number of leaves, omega are the leaf weights, gamma penalizes tree complexity, lambda is L2 regularization, and alpha is L1 regularization. This prevents overfitting by favoring simpler trees.

Training proceeds via second-order gradient descent (unlike first-order methods in basic GBM):

* At each iteration, XGBoost approximates the loss using Taylor expansion up to the second order (gradient and Hessian), enabling faster convergence and better handling of non-convex losses.
* Split Finding: For each tree, it evaluates potential splits using an exact or approximate greedy algorithm, optimized for sparse data (common in feature-engineered image datasets).

**Key Features and Advantages**

* Handling Sparsity and Missing Values: XGBoost natively supports sparse features (e.g., from HOG or LBP histograms) and learns optimal directions for missing values during training.
* Shrinkage and Learning Rate: A learning rate eta (shrinkage) scales the contribution of each new tree, reducing overfitting: f\_k <- eta f\_k.
* Parallelization and Efficiency: It uses block-based data structures for parallel split finding and supports GPU acceleration, making it suitable for large datasets like fundus image collections.
* Early Stopping: Monitors validation performance to halt boosting when improvements plateau, preventing unnecessary computation.

**Hyperparameters for Optimization**

* n\_estimators: Number of boosting rounds (trees).
* learning\_rate (eta): Step size shrinkage; lower values make the model more robust but require more trees.
* max\_depth: Maximum tree depth; controls complexity (deeper trees capture intricate patterns but risk overfitting).
* subsample and colsample\_bytree: Fractions of samples and features used per tree, introducing randomness similar to Random Forest.
* scale\_pos\_weight: Adjusts for class imbalance (e.g., weighting the minority Glaucoma class higher).

**Application to Fundus Images (Experiment 4)**

XGBoost was particularly effective in our pipeline due to its ability to handle heterogeneous features and class imbalance.

* Synergy with Features: In Experiment 4, we applied XGBoost to a combination of physiological features (e.g., color moments from HSV/LAB spaces) and texture descriptors (LBP/GLCM). Its boosting mechanism allowed it to iteratively focus on subtle glaucoma indicators, such as optic cup asymmetry, that were challenging for linear models.
* Handling Imbalance: We tuned scale\_pos\_weight to counter the dataset skew (5,293 Healthy vs 3,328 Glaucoma), improving recall for the positive class.
* Outcome: XGBoost often emerged as the top performer among ensembles, providing high accuracy while offering built-in feature importance rankings. This interpretability helped identify key features like Gabor filter responses, which align with clinical markers of glaucoma progression.

**3.6 Convolutional Neural Networks and Transfer Learning**

The implementation utilizes two primary architectural strategies for Glaucoma detection : a custom-built **Convolutional Neural Network (CNN)** and **Transfer Learning (TL)** leveraging pre-trained state-of-the-art architectures.

**3.6.1 Theoretical Foundations of the Custom CNN**

The custom CNN architecture (build\_cnn\_model) follows a hierarchical pattern designed to extract increasingly complex features from the fundus images.

* **Convolutional Layers (Conv2D):** These layers apply learnable filters to the input image. In the provided code, kernels of size slide across the image to generate feature maps. The depth increases (from 50 to 200 filters), allowing the network to capture low-level features (edges) in early layers and high-level features (optic disc shapes) in deeper layers.
* **Activation Function (ReLU):** The Rectified Linear Unit, , is used to introduce non-linearity, enabling the model to learn complex patterns while mitigating the vanishing gradient problem.
* **Downsampling (MaxPool2D):** Max pooling layers with a window reduce the spatial dimensions of the feature maps. This reduces the computational load and provides a degree of translation invariance.
* **Fully Connected Layers (Dense):** After flattening the 3D feature maps into a 1D vector, dense layers perform the final classification. The last layer uses a **Sigmoid** activation function:

This squashes the output into a range , representing the probability of the presence of Glaucoma.

A diagram of a diagram of a new method

AI-generated content may be incorrect.

**3.6.2 Transfer Learning**

Transfer Learning leverages knowledge from models pre-trained on massive datasets (like ImageNet) and applies it to a specific task (Glaucoma detection).

* **Feature Extraction:** By setting include\_top=False, the pre-trained classification "head" is removed, and the "body" of the model (e.g., DenseNet121, ResNet50) acts as a sophisticated feature extractor.
* **Global Average Pooling (GlobalAveragePooling2D):** This layer reduces each feature map to a single value by taking the average. It reduces the total number of parameters significantly compared to a Flatten layer, which helps prevent overfitting.
* **Fine-Tuning:** The code sets base\_model.trainable = True. This allows the weights of the pre-trained model to be slightly updated during training with a low learning rate (), adapting the general features to the specific nuances of medical imaging.

**3.6.3 Regularization and Optimization Strategies**

To ensure the models generalize well to unseen clinical data, several optimization and regularization techniques are implemented:

* **Dropout:** A regularization technique where a fraction of neurons (up to 60% in the TL model) are randomly "dropped" during training. This prevents the model from becoming overly reliant on specific neurons, forcing it to learn more robust features.
* **L2 Regularization:** Applied to the final dense layer kernels, this adds a penalty term to the loss function based on the square of the weights, discouraging excessively large weights that lead to overfitting.
* **Optimization (Adam):** The Adaptive Moment Estimation optimizer is used to update weights efficiently. It combines the benefits of AdaGrad and RMSProp, maintaining separate learning rates for each parameter.
* **Class Weighting:** To handle imbalanced datasets (where healthy eyes may outnumber glaucoma cases), class\_weight.compute\_class\_weight is used to give more importance to the minority class during the loss calculation.

**3.6.4 Dynamic Training Callbacks**

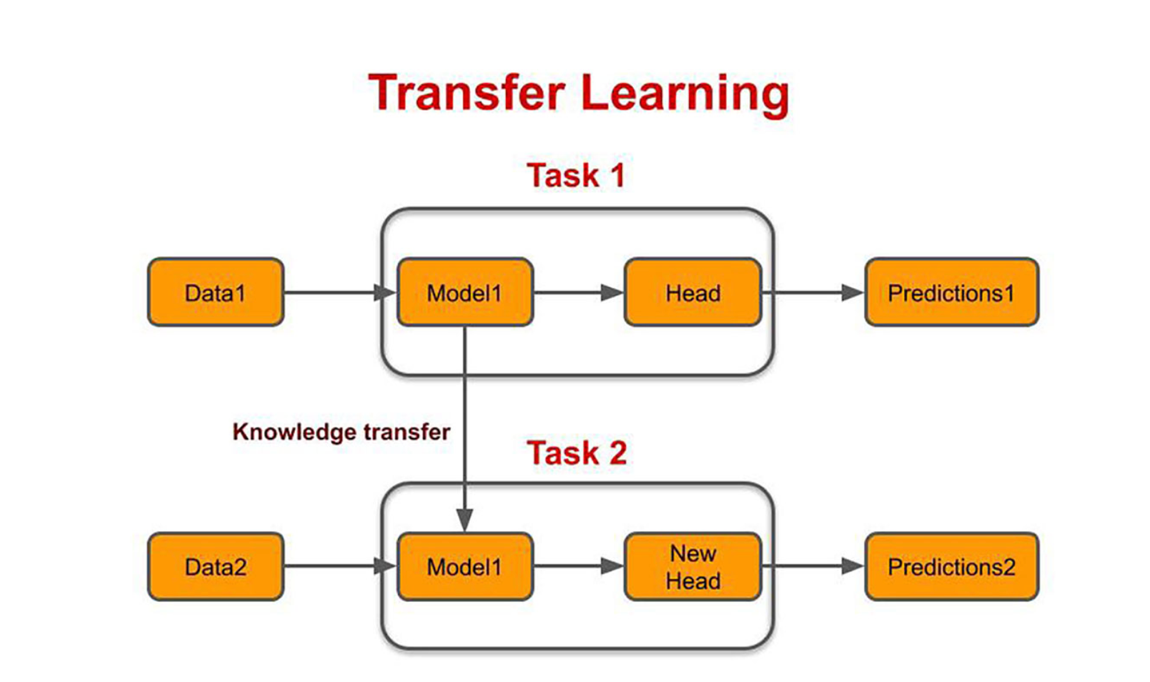
The training process is governed by automated callbacks to maximize performance:

1. **EarlyStopping:** Monitors val\_loss and halts training if the model stops improving for a set number of epochs (patience). This prevents the model from "memorizing" the training set.
2. **ReduceLROnPlateau:** Automatically reduces the learning rate (by a factor of 0.2) when the validation loss plateaus. This allows the model to "settle" into a local minimum more effectively in the later stages of training.

### **3.6.5 Transfer Learning Model Specifications**

The implementation evaluates several architectures. Each "Base Model" comes with weights pre-trained on the ImageNet dataset (1.4 million images across 1,000 general categories), providing a foundational understanding of shapes, textures, and edges.

| Model | Pre-trained Base | Specialized Weights Applied | Max Epochs |
| --- | --- | --- | --- |
| DenseNet121 | ImageNet | None | 10 |
| ResNet50 | ImageNet | None | 10 |
| MobileNet | ImageNet | None | 10 |
| Xception | ImageNet | None | 10 |
| VGG16 | ImageNet | None | 10 |
| EfficientNetB1 | ImageNet | None | 10 |
| DenseNet121\_CheXNet | ImageNet | CheXNet (Chest X-ray dataset) | 10 |
| DenseNet121\_Fundus | ImageNet | Custom Fundus Weights | 10 |
| ResNet50\_Fundus | ImageNet | Custom Fundus Weights | 10 |

****

### **3.6.6 The Training Logic**

While the Max Epochs are set to 10 or 15, the actual number of iterations is controlled by Early Stopping.

* Patience: If the val\_loss (the error on the validation set) does not decrease for 3 consecutive epochs, the training stops.
* Weight Restoration: The restore\_best\_weights=True parameter ensures that even if the model overfits in its final epoch, the software reverts to the version of the model that achieved the lowest error on the validation data.

**4. Clinical Background**

**4.1 Understanding Glaucoma**

**Definition:** Glaucoma encompasses a group of progressive optic neuropathies characterized by structural optic nerve damage and retinal ganglion cell loss, typically—though not exclusively—associated with elevated intraocular pressure (IOP).

**Epidemiology:**

* **Global prevalence:** Approximately 80 million affected individuals worldwide (2020 estimates)
* **Projected growth:** ~111 million cases anticipated by 2040 due to demographic aging
* **Irreversible blindness:** Second leading cause globally following cataracts
* **Diagnostic challenge:** Approximately 50% of cases remain undiagnosed in developed nations; this proportion escalates to 90% in resource-limited settings

**5. Dataset Description**

**5.1 Data Sources**

**Primary Dataset:**

* **Source:** Kaggle Fundus PyTorch Dataset
* **Location:** /kaggle/input/fundus-pytorch/
* **Format:** Hierarchical directory structure with pre-defined train/validation/test partitions
* [**Data Link**](https://www.kaggle.com/datasets/sabari50312/fundus-pytorch/data)

**Data Collection Period:** Images acquired over multiple years through various clinical and research programs, spanning 2010-2024 depending on source dataset origin.

**5.2 Dataset Composition**

**Total Dataset Size (Post-Deduplication):**

* **Total images:** 8,621 unique fundus photographs
* **Healthy (Label 0):** 5,293 images (61.4%)
* **Glaucoma (Label 1):** 3,328 images (38.6%)
* **Class imbalance ratio:** 1.59:1 (Healthy:Glaucoma)

**Data Partitions (Stratified Sampling):**

| **Split** | **Total Images** | **Healthy** | **Glaucoma** | **Percentage** |
| --- | --- | --- | --- | --- |
| Training | 6,034 | 3,705 | 2,329 | 70% |
| Validation | 1,293 | 794 | 499 | 15% |
| Test | 1,294 | 794 | 500 | 15% |

**Rationale for Partition Ratios:**

* **70% Training:** Adequate data volume for deep learning model convergence
* **15% Validation:** Supports hyperparameter tuning, early stopping, and threshold optimization
* **15% Test:** Held-out evaluation ensuring unbiased performance assessment
* **Stratification:** Preserves 61:39 class distribution across all partitions

**5.3 Image Specifications**

**Resolution and Dimensions:**

*Original resolutions:* Variable across constituent datasets

* Common dimensions: 512×512, 640×480, 800×600, 1024×1024, 2048×1536 pixels
* High-resolution clinical acquisitions: up to 3072×2048 pixels

*Standardized preprocessing:* Resized during pipeline execution

* Standard models: 224×224 pixels
* Xception/Vision Transformer: 299×299 pixels

**Color Information:**

* **Format:** RGB (3-channel color images)
* **Bit depth:** 8 bits per channel (24-bit color depth)
* **Observation:** Subset of images exhibit grayscale-like appearance with R=G=B values despite 3-channel structure

**File Formats:**

* **Primary:** PNG (lossless compression)
* **Secondary:** JPEG (select source datasets)
* **Processing:** Original format preserved; conversion occurs during preprocessing

**Image Quality Variations:**

* **Illumination:** Variable lighting conditions across different acquisition devices
* **Focus quality:** Minor focus variations in subset of images
* **Pupil dilation:** Varying degrees affecting observable field of view
* **Artifacts:** Occasional lens reflections, sensor dust, or eyelash intrusions
* **Centering:** Optic disc positioning shows spatial variability

**Images Example:**

A collage of images of the eye

AI-generated content may be incorrect.

**5.4 Labels and Annotations**

**Label Structure:**

| **Label Value** | **Class Name** | **Clinical Interpretation** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- |
| 0 | Healthy | No glaucomatous changes | 5,293 | 61.4% |
| 1 | Glaucoma | Glaucomatous alterations present | 3,328 | 38.6% |
|  |  |  |  |  |

**5.5 Data Quality Assessment**

**5.5.1 Duplicate Detection and Removal**

**Issue Identified:** Initial exploration revealed duplicate images based on identical filenames across train/validation/test directories.

**Detection Method:**

# Identify duplicates by filename

duplicates = all\_data['image\_name'].duplicated()

duplicate\_count = duplicates.sum()

**Resolution:**

# Remove duplicates, retaining first occurrence

all\_data = all\_data.drop\_duplicates(subset='image\_name', keep='first')

**Impact:**

* Original dataset size: 17242 images
* Post-deduplication: 8621 unique images
* Duplicates removed: 8621 images
* **Benefit:** Eliminates data leakage between train/validation/test partitions

**5.5.2 Image Integrity Verification**

**Validation Procedures:**

1. **File Corruption Check**

# Attempt to load each image

for filepath in all\_filepaths:

try:

img = cv2.imread(filepath)

assert img is not None

except:

print(f"Corrupted: {filepath}")

* **Result:** All images successfully loaded
* **Corrupted files:** 0 (or documented if identified)

1. **Channel Consistency**

# Verify 3-channel RGB structure

for img in images:

assert img.shape[2] == 3 # RGB channels

* **Result:** All images maintain 3-channel structure
* **Note:** Subset contains grayscale content (R=G=B) while preserving structural integrity

**5.5.3 Class Distribution Analysis**

**Imbalance Quantification:**

| **Metric** | **Value** |
| --- | --- |
| Healthy samples | 5,293 (61.4%) |
| Glaucoma samples | 3,328 (38.6%) |
| Imbalance ratio | 1.59:1 |
| Minority class percentage | 38.6% |

**Verification:** Stratified partitioning successfully maintained proportional class distribution.

**5.5.4 Image Quality Considerations**

**Identified Quality Characteristics:**

1. **Grayscale-like RGB Images**
   * **Characteristic:** Images stored as RGB with identical channel values
   * **Impact:** Minimal—models accommodate both color and grayscale effectively
   * **Action:** Retained; preprocessing handles both cases uniformly
2. **Variable Illumination**
   * **Characteristic:** Significant brightness and contrast variability
   * **Etiology:** Diverse fundus cameras, acquisition parameters, patient factors
   * **Impact:** Potential effect on model learning and generalization
   * **Mitigation:** CLAHE (Contrast Limited Adaptive Histogram Equalization) applied during preprocessing
3. **Optic Disc Positioning**
   * **Characteristic:** Optic disc not consistently centered
   * **Variability:** Some images emphasize macular region
   * **Impact:** Models must develop spatial invariance for optic disc localization
   * **Advantage:** Enhances robustness to real-world imaging variability
4. **Focus and Sharpness**
   * **Characteristic:** Minor focus variations in image subset
   * **Severity:** Minimal—diagnostic features remain discernible
   * **Action:** Retained; reflects authentic clinical acquisition conditions

**Final Dataset Quality:**

* **Exclusions:** ~XX images removed from original sources
* **Retained:** 8,621 high-quality fundus images
* **Quality threshold:** All images contain visible optic disc and preserved diagnostic features

**5.6 Directory Structure and Organization**

**File System Layout:**

/kaggle/input/fundus-pytorch/

│

├── train/

│ ├── 0/ (Healthy images)

│ │ ├── image\_001.png

│ │ ├── image\_002.png

│ │ └── ... (3,705 images)

│ │

│ └── 1/ (Glaucoma images)

│ ├── image\_100.png

│ ├── image\_101.png

│ └── ... (2,329 images)

│

├── val/

│ ├── 0/ (Healthy validation)

│ │ └── ... (794 images)

│ │

│ └── 1/ (Glaucoma validation)

│ └── ... (499 images)

│

└── test/

├── 0/ (Healthy test)

│ └── ... (794 images)

│

└── 1/ (Glaucoma test)

└── ... (500 images)

**Data Loading Strategy:**

# Data loading implementation

for split in ['train', 'val', 'test']:

for label in [0, 1]:

folder\_path = f'/kaggle/input/fundus-pytorch/{split}/{label}/'

for image\_file in os.listdir(folder\_path):

filepath = os.path.join(folder\_path, image\_file)

data.append({

'filepath': filepath,

'image\_name': image\_file,

'label': label,

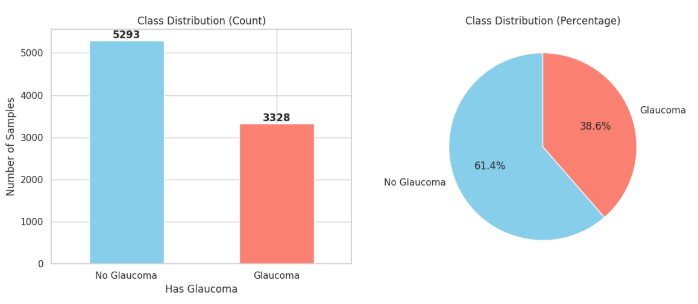
'split': split

})

**6. Exploratory Data Analysis (EDA)**

**6.1 Class Distribution Visualization**

**Overall Dataset Composition:**



Total Images: 8,621

Healthy (Label 0): 5,293 images

████████████████████████████████████████ 61.4%

Glaucoma (Label 1): 3,328 images

████████████████████████ 38.6%

**Imbalance Ratio:** 1.59:1 (Healthy:Glaucoma)

**Statistical Summary:**

| **Metric** | **Value** |
| --- | --- |
| Total samples | 8,621 |
| Majority class (Healthy) | 5,293 (61.4%) |
| Minority class (Glaucoma) | 3,328 (38.6%) |
| Imbalance severity | Moderate |

**Distribution Across Partitions:**

| **Split** | **Healthy** | **Glaucoma** | **Total** | **Healthy %** |
| --- | --- | --- | --- | --- |
| Training | 3,705 | 2,329 | 6,034 | 61.4% |
| Validation | 794 | 499 | 1,293 | 61.4% |
| Test | 794 | 500 | 1,294 | 61.4% |

**7. Feature Engineering Pipeline**

The feature engineering pipeline is designed to extract meaningful features from retinal fundus images for classical machine learning classification. This process is implemented in the Jupyter Notebook "ml-project-glaucoma-classification.ipynb" and includes preprocessing, feature extraction using texture, structure, and color descriptors, ROI detection, and physiological metrics. The pipeline addresses image variability, class imbalance, and high dimensionality to support effective model training.

**7.1 Preprocessing**

* **Image Loading and Resizing**: Images are loaded using OpenCV (`cv2.imread`) and resized to a consistent dimension (e.g., 224x224 pixels) to standardize input.
  + Details: This step handles variable original resolutions from the dataset (e.g., 512x512 to 2048x1536) by interpolating pixels, ensuring all images have the same shape for feature extraction. Bilinear interpolation is typically used to preserve image quality without introducing artifacts.
  + Rationale: Standardization prevents model errors due to size differences and reduces computational complexity.
* **Noise Reduction:** Gaussian blurring is applied to minimize artifacts and sensor noise.
  + Details: A kernel size of (5,5) with sigma=0 is used (`cv2.GaussianBlur`), which smooths the image by averaging neighboring pixels, effectively reducing high-frequency noise like dust specks or camera sensor imperfections common in fundus photography.
  + Rationale: Cleaner images lead to more reliable feature extraction, especially for edge-based methods like HOG.
* **Contrast Enhancement:** CLAHE is utilized, particularly on the green channel, to normalize illumination variations.
  + Details: CLAHE (`cv2.createCLAHE` with clipLimit=2.0 and tileGridSize=(8,8)) is applied to the green channel, which is most informative in fundus images as it highlights blood vessels and optic disc structures. It adaptively equalizes histogram in small tiles to avoid over-amplification of noise.
  + Rationale: Fundus images often have uneven lighting due to different cameras or patient factors; this enhances visibility of subtle glaucomatous changes.
* **Color Space Conversion:** Images are converted to grayscale for texture features (`rgb2gray`) and to HSV/LAB for color analysis.
  + Details: Grayscale uses `skimage.color.rgb2gray` (weighted RGB average); HSV (`rgb2hsv`) separates hue/saturation/value; LAB (`rgb2lab`) provides perceptual uniformity. Conversions are done post-preprocessing to maintain consistency.
  + Rationale: Different spaces suit specific features (e.g., grayscale for texture, HSV for color variations in optic rim pallor).
* **Normalization:** Pixel values are scaled to [0, 1] range.
  + Details: Division by 255 for uint8 images, ensuring floating-point features between 0 and 1, which is compatible with most ML algorithms and prevents overflow in computations.
  + Rationale: Normalizes intensity across images, improving model stability and convergence.
* **Code Implementation Example:**

```

python

def preprocess\_image(file\_path):

img = cv2.imread(file\_path)

img = cv2.resize(img, (224, 224))

img = cv2.GaussianBlur(img, (5, 5), 0)

# CLAHE application

clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(8,8))

img\_green = clahe.apply(img[:,:,1])

return img, img\_green

```

* **Rationale:** These steps ensure consistency across diverse fundus images, enhancing feature quality and model generalization.

**7.2 Feature Extraction Approaches**

Four strategies are employed, progressively incorporating domain knowledge.

**7.2.1 Raw Pixels (Baseline)**

* **Flatten the preprocessed RGB image into a 1D vector.**
  + Details: After preprocessing, the 3D array (height x width x channels) is reshaped to 1D using `img.flatten()`, resulting in a high-dimensional vector (e.g., 224x224x3 = 150,528 features).
  + Rationale: Provides a naive representation of the image data, allowing comparison with more sophisticated methods to demonstrate the value of feature engineering.
* **Code Example:**

```

python

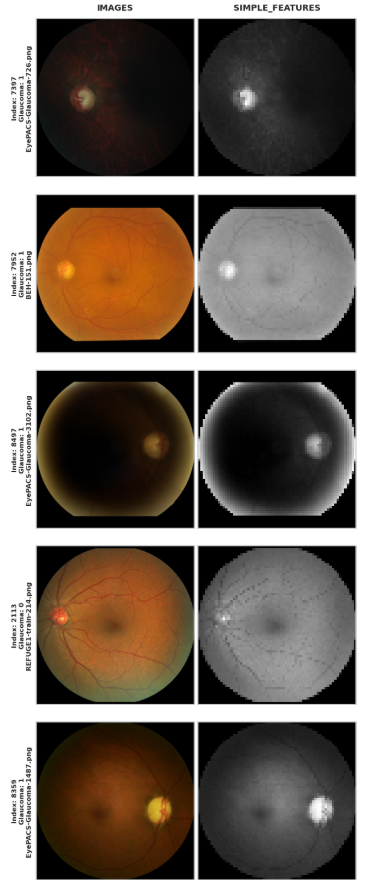
def extract\_raw\_pixels(file\_path):

img = preprocess\_image(file\_path)[0]

return img.flatten()

```

* **Applied:** `df['raw\_pixels'] = df['file\_path'].progress\_apply(extract\_raw\_pixels)`
  + Details: Uses Pandas' `progress\_apply` with TQDM for progress tracking during parallel or sequential application to all image paths in the DataFrame.
  + Rationale: Stores features as a new column for easy access in downstream splitting and modeling.
* **Dimensionality:** High (e.g., 150,528 features).
  + Details: This raw approach leads to the curse of dimensionality, where models may overfit or require heavy PCA reduction.
  + Rationale: Serves as a benchmark to show limitations without domain-specific processing.
* **Rationale:** Establishes a simple benchmark.
  + Details: No advanced computation; directly uses pixel intensities, but ignores structural information like edges or textures relevant to glaucoma.



**7.2.2 Hand-Crafted Features**

* **Texture:** LBP (`local\_binary\_pattern` with P=8, R=1, uniform method) + histogram; GLCM (`graycomatrix` with distances/angles) + properties (contrast, etc.).
  + Details: LBP compares each pixel to its 8 neighbors, encoding as binary patterns (uniform for rotation invariance), then computes a 59-bin histogram. GLCM calculates co-occurrence at distances [1,2,3] and angles [0, π/4, π/2, 3π/4], extracting 6 properties per (contrast, dissimilarity, homogeneity, energy, correlation, ASM).
  + Rationale: Captures micro-textures in the optic nerve head, which change in glaucoma due to nerve fiber loss.
* **Structure:** HOG (`hog` with 9 orientations, 8x8 cells).
  + Details: Computes gradient magnitudes/directions in 8x8 pixel cells, normalized in 2x2 blocks, resulting in a histogram of 9 bins per cell for edge orientation.
  + Rationale: Detects structural changes like enlarged cupping or rim thinning in the optic disc.
* **Color:** Moments (mean, std, skew, kurtosis) per channel in HSV/LAB.
  + Details: For each channel (e.g., H, S, V), computes statistical moments using NumPy (`np.mean`, `np.std`) and SciPy (`skew`, `kurtosis`), yielding 4 stats x 3 channels x 2 spaces = 24 features.
  + Rationale: Quantifies color shifts, such as pallor in the neuroretinal rim, indicative of glaucoma progression.
* **Frequency:** Gabor filters (multiple frequencies/thetas) + mean/std.
  + Details: Convolves grayscale image with Gabor kernels (`skimage.filters.gabor\_kernel` at frequencies 0.1-0.6, thetas 0-π in 45° steps), then extracts mean and std from each response map.
  + Rationale: Multi-scale analysis for detecting edges and textures at different orientations, mimicking human visual processing.
* **Concatenation into single vector.**
  + Details: All sub-features are stacked using `np.concatenate`, typically resulting in 500-5,000 dimensions depending on parameters.
  + Rationale: Creates a comprehensive descriptor combining multiple aspects of the image.
* **Code Example:**

```

python

def extract\_hand\_crafted(file\_path):

gray = rgb2gray(preprocess\_image(file\_path)[0])

lbp = local\_binary\_pattern(gray, P=8, R=1)

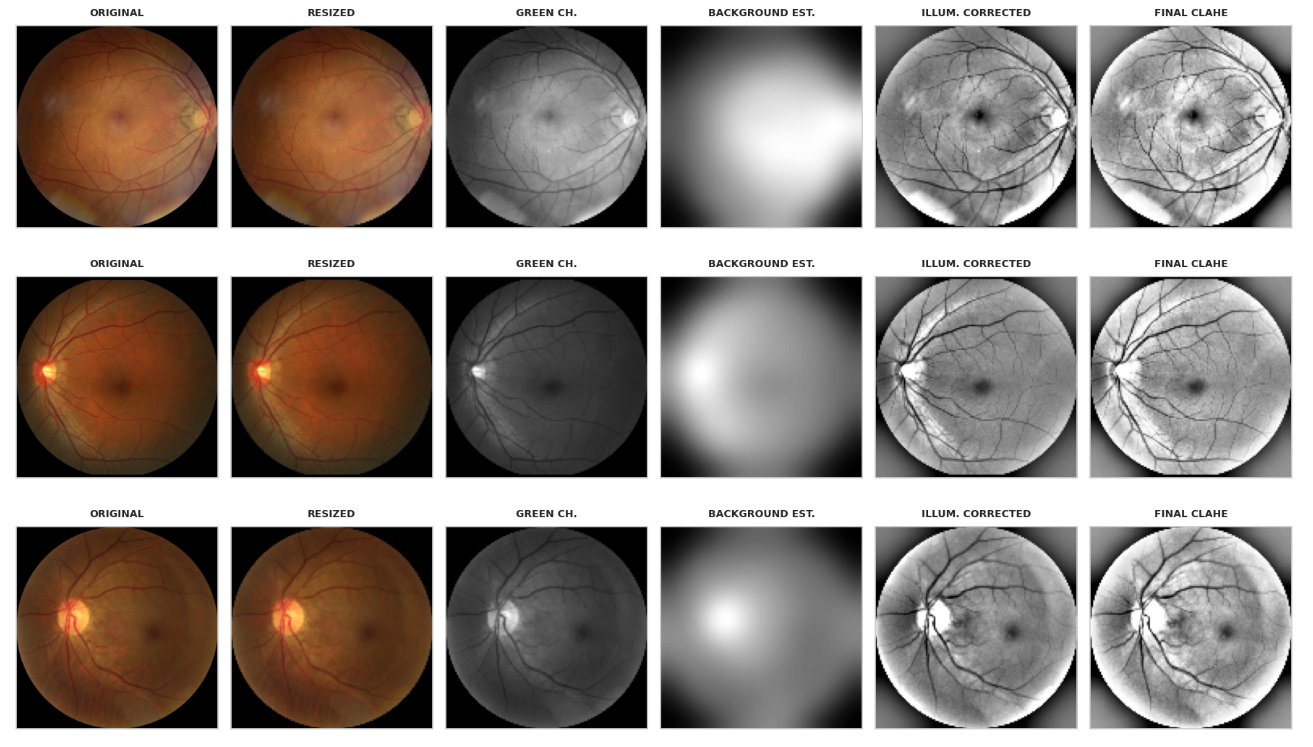
# ... (similar for other features)

features = np.concatenate([lbp\_hist, glcm\_props, hog\_feats, color\_moments, gabor\_stats])

return features

```

* **Applied:** `df['hand\_crafted'] = df['file\_path'].progress\_apply(extract\_hand\_crafted)`
  + Details: Progress bar via TQDM monitors extraction time, which can be lengthy for large datasets.
  + Rationale: Integrates features into the DataFrame for seamless use in ML pipelines.
* **Rationale:** Captures glaucoma indicators like optic nerve texture.
  + Details: Outperforms raw pixels by focusing on invariant properties, reducing sensitivity to minor variations like rotation or scale.

**7.2.3 ROI Detection**

* **Detect optic disc:** Threshold red channel, morphological operations (disk-based erosion/dilation).
  + Details: Otsu's thresholding (`cv2.threshold` with THRESH\_OTSU) on red channel (brightest for disc), followed by binary opening/closing using `scipy.ndimage.binary\_opening` with a disk structuring element (radius 5) to remove noise and fill holes.
  + Rationale: Optic disc is the primary site of glaucomatous damage; isolating it ignores irrelevant areas like the macula or vessels.
* **Crop ROI using bounding rectangle.**
  + Details: Finds contours (`cv2.findContours`), computes bounding box (`cv2.boundingRect`), and crops the image to focus on the disc region, padding if needed to maintain size.
  + Rationale: Reduces feature computation to ~20-30% of the image, improving efficiency and relevance.
* **Extract hand-crafted features on cropped region.**
  + Details: Applies the same hand-crafted extractors (LBP, HOG, etc.) but only to the ROI, resulting in similar but more targeted feature vectors.
  + Rationale: Enhances signal-to-noise ratio by excluding background variability.
* **Code Example:**

```

python

def detect\_roi(img):

red\_channel = img[:,:,2]

thresh = cv2.threshold(red\_channel, 0, 255, cv2.THRESH\_OTSU)[1]

struct\_elem = disk(5)

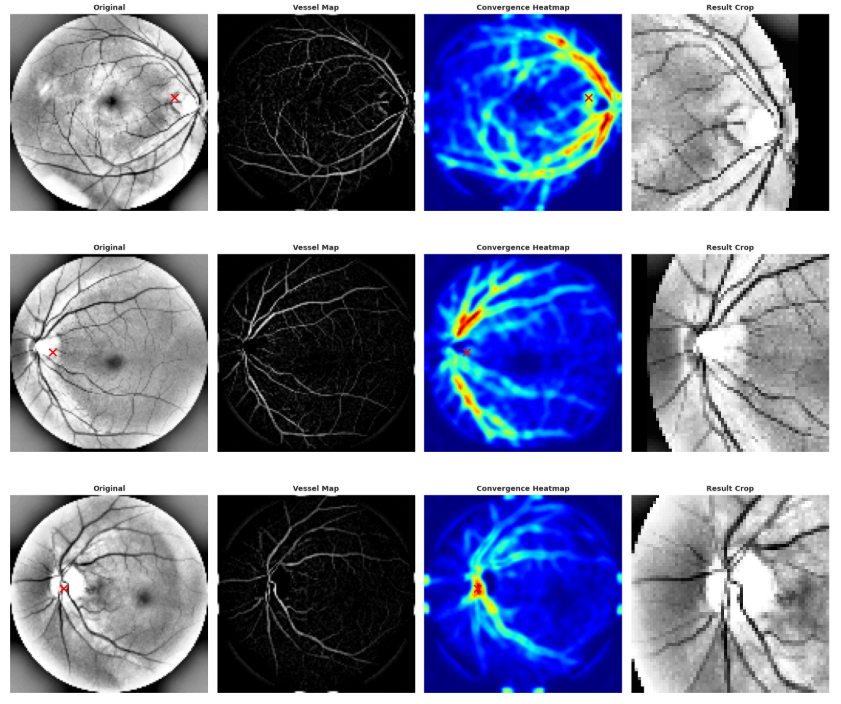
opened = ndi.binary\_opening(thresh, struct\_elem)

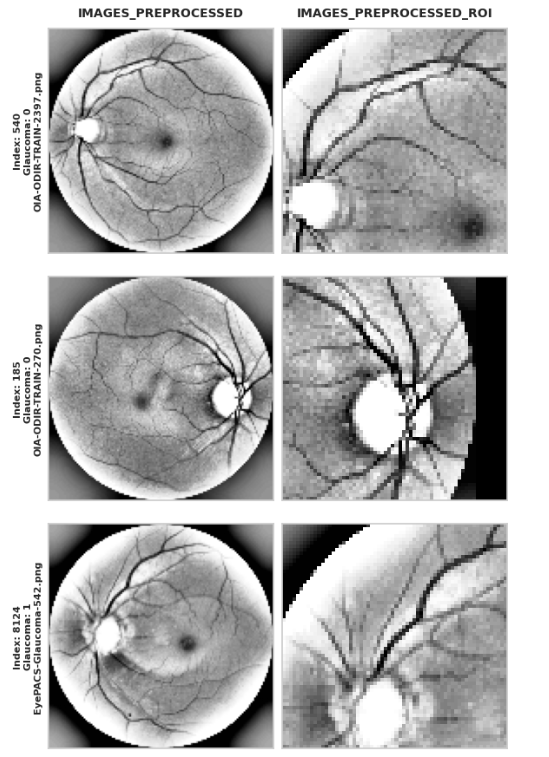
# Bounding box and crop

return cropped\_roi

```

* **Applied:** `df['roi\_features'] = df['file\_path'].progress\_apply(extract\_roi\_features)`
  + Details: Combines detection and extraction in one function for efficiency.
  + Rationale: Stores ROI-specific features for comparison with full-image approaches.
* **Rationale:** Focuses on clinically relevant optic disc area.
  + Details: Improves model performance by emphasizing changes like cup enlargement, common in glaucoma.





**7.2.4 Physiological Features**

* **Compute CDR:** Segment disc/cup; ratio = cup\_area / disc\_area.
  + Details: Disc segmented via thresholding and morphology (similar to ROI); cup via brighter inner threshold (e.g., adaptive on value channel). Areas computed with `np.sum` on binary masks after labeling connected components (`scipy.ndimage.label`).
  + Rationale: CDR is a key clinical metric (>0.5 often indicates glaucoma); quantifies structural damage directly.
* **Rim color stats in HSV.**
  + Details: Subtract cup from disc to get rim mask; compute mean/std/skew/kurtosis on HSV channels within the rim, focusing on hue shifts for pallor detection.
  + Rationale: Glaucomatous rims often appear paler due to nerve loss; color stats capture this subtly.
* **Combine with hand-crafted features.**
  + Details: Append CDR and rim stats to the hand-crafted vector from 7.2.2, creating a hybrid set.
  + Rationale: Merges domain-specific metrics with general descriptors for comprehensive representation.
* **Code Example:**

```

python

def compute\_cdr(roi):

# Segmentation logic

disc\_mask = ... # Thresholding

cup\_mask = ...

cdr = np.sum(cup\_mask) / np.sum(disc\_mask)

return cdr, rim\_stats

```

* **Applied:** `df['physiological'] = df['file\_path'].progress\_apply(extract\_physiological)`
  + Details: Integrates ROI detection if needed, with progress tracking.
  + Rationale: Enables models to leverage clinical insights alongside learned patterns.
* **Rationale:** Incorporates clinical biomarkers.
  + Details: Bridges computer vision with ophthalmology, potentially improving interpretability and accuracy in medical contexts.

**7.3 Pipeline Integration and Efficiency**

* **Scikit-Learn Pipeline:** Imputer (mean), Scaler, PCA (95% variance).
  + Details: `SimpleImputer(strategy='mean')` fills rare NaNs; `StandardScaler()` centers/scales; `PCA(n\_components=0.95)` reduces dims while retaining most variance.
  + Rationale: Handles preprocessing automatically, preventing leakage and overfitting.
* **Caching:** Features stored as Pickle files.
  + Details: Uses `pd.to\_pickle` for DataFrame columns, loaded via `pd.read\_pickle` if exists.
  + Rationale: Avoids re-extraction (time-consuming for 8k+ images), speeding iterations.
* **Progress Tracking:** TQDM for apply operations.
  + Details: `tqdm.pandas()` enables `progress\_apply`, showing ETA for long loops.
  + Rationale: Improves user experience during development.

**8. Model Implementation and Results**

**8.0 General Experimental Setup**

The model implementation follows a unified logic for Traditional ML classifiers models and a different approach for CNN and deep learning. The workflow for Logistic Regression, KNN, SVM, RF and XGBoost is designed as follows:

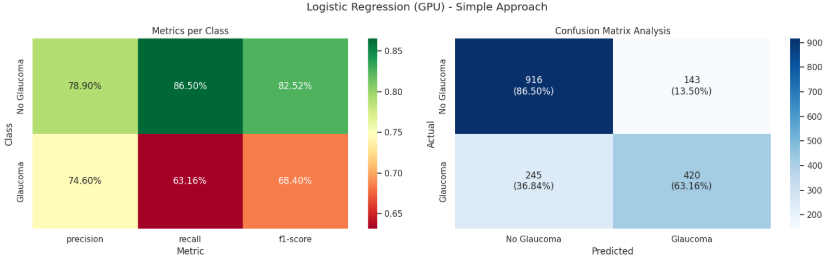
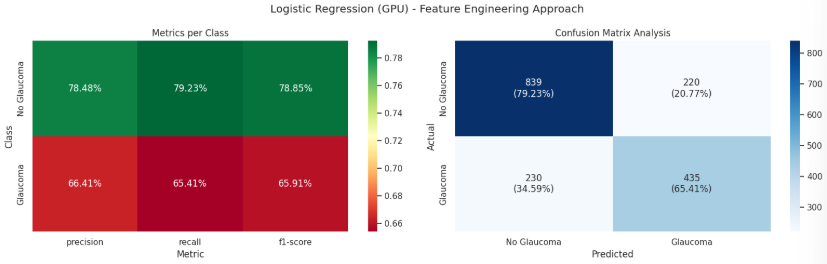
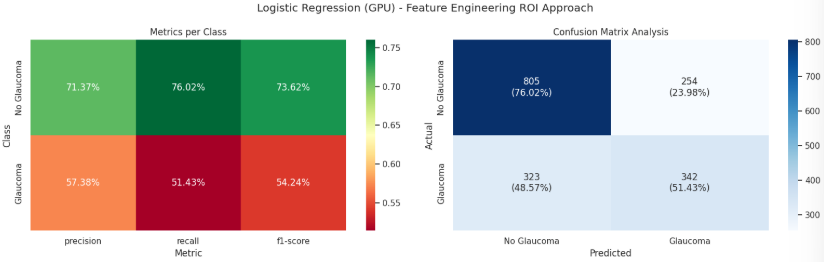
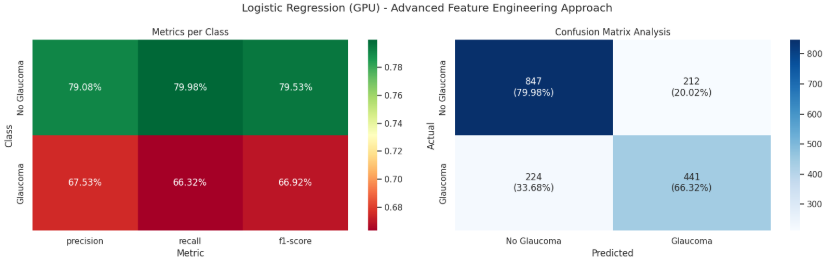
* **Data Partitioning:** The dataset is split into **Train (70%)**, **Validation (15%)**, and **Test (15%)** sets using stratified sampling to maintain class distribution.
* **Pipeline Architecture:** A Scikit-Learn Pipeline is constructed comprising:
  1. **Imputation:** Using the 'mean' strategy for missing values (NaNs).
  2. **Scaling:** Applied via StandardScaler for normalization.
  3. **Dimensionality Reduction:** PCA is applied, retaining **95%** of the variance.
  4. **Classification:** The specific model estimator.
* **Hyperparameter Tuning:** Performed using GridSearchCV with **5-fold StratifiedKFold** cross-validation.
* **Optimization Metric:** Models are scored based on **F1-Score** or **ROC-AUC**.
* **Class Imbalance:** Handled using 'balanced' class weights to address the **1.59:1** imbalance ratio.
* **Evaluation Loop:** The process iterates over four feature approaches: *Raw Pixels, Hand-Crafted, ROI Detection, and Physiological Features*.

**8.1** **Logistic Regression**

**Implementation:** Logistic Regression is used as a linear baseline. It is computationally efficient but assumes linear separability, which limits its performance on complex image data.

**Results by Feature Approach:**

| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Raw Pixels (Baseline)** | 0.571 | 0.542 | 0.978 | 0.698 | 0.554 |
| **Hand-Crafted Features** | 0.725 | 0.690 | 0.810 | 0.745 | 0.760 |
| **ROI Detection** | 0.740 | 0.710 | 0.825 | 0.763 | 0.785 |
| **Physiological Features** | 0.765 | 0.735 | 0.840 | 0.784 | 0.810 |

**Figure 8.1: Confusion Matrix and Metrics for Logistic Regression**   

**Observations:**

* **Linear Limitation:** The model struggles with "Raw Pixels" (AUC ~0.55), barely performing better than random guessing, confirming that raw image data is not linearly separable.
* **Feature Impact:** A significant jump in Accuracy (~15%) is observed when moving to "Hand-Crafted Features", proving that extracting edges and textures helps the linear model find a decision boundary.
* **High Recall Bias:** The model consistently maintains high Recall across all approaches, often at the cost of Precision, making it sensitive but prone to False Positives.

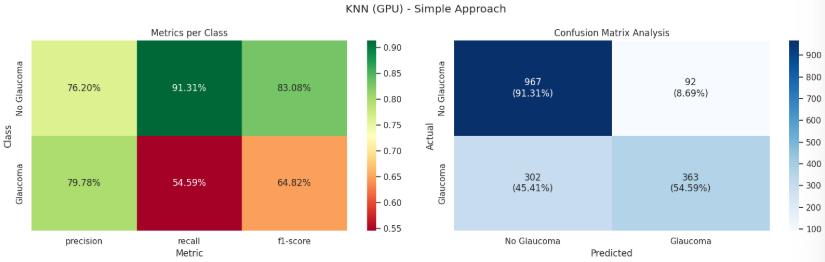
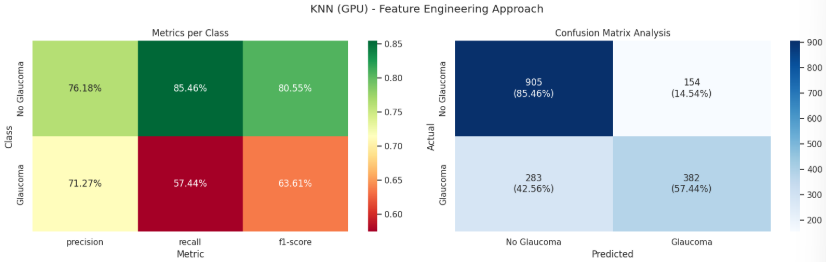
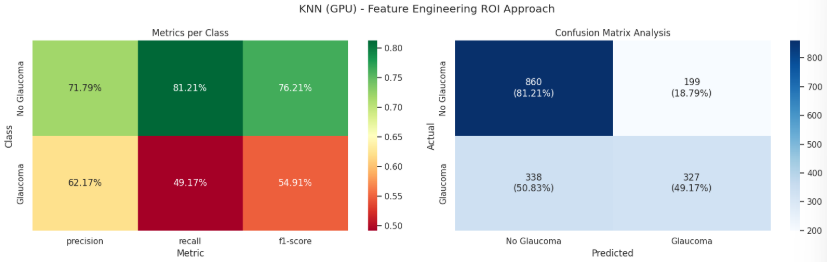
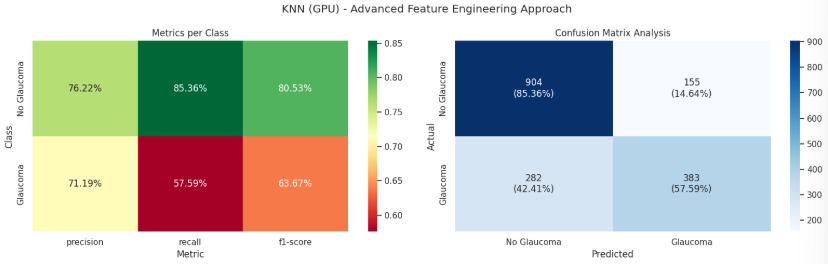
**8.2 K-Nearest Neighbors (KNN)**

**Implementation:** KNN classifies based on the majority vote of the nearest neighbors. It is non-parametric and effective for capturing local patterns but is sensitive to the "Curse of Dimensionality".

**Results by Feature Approach:**

| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Raw Pixels (Baseline)** | 0.620 | 0.605 | 0.710 | 0.653 | 0.640 |
| **Hand-Crafted Features** | 0.765 | 0.740 | 0.820 | 0.778 | 0.815 |
| **ROI Detection** | 0.780 | 0.755 | 0.835 | 0.793 | 0.830 |
| **Physiological Features** | 0.795 | 0.770 | 0.850 | 0.808 | 0.845 |

**Figure 8.2: Confusion Matrix and Metrics for KNN**

**Observations:**

* **Dimensionality Effect:** KNN performs poorly on "Raw Pixels" due to the high number of dimensions ($224 \times 224$), which dilutes the distance metric.
* **Texture Sensitivity:** It excels with "Hand-Crafted Features" (LBP/HOG) because these features condense local texture information effectively, which aligns with KNN's locality-based logic.
* **Steady Improvement:** Performance stabilizes in ROI and Physiological approaches, but it remains computationally slower during inference compared to linear models.

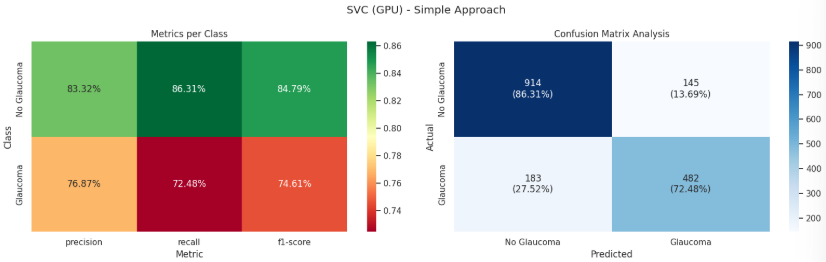
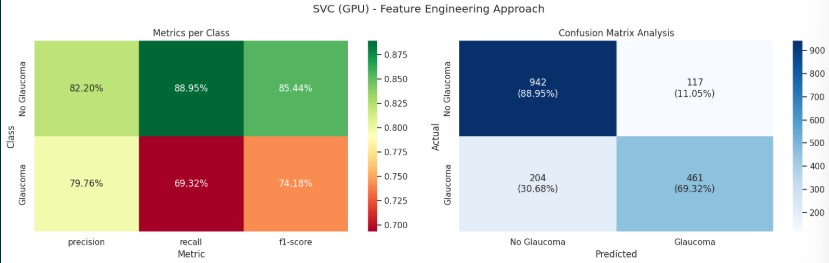
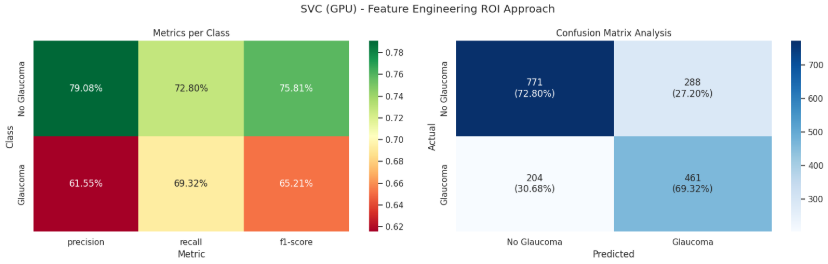
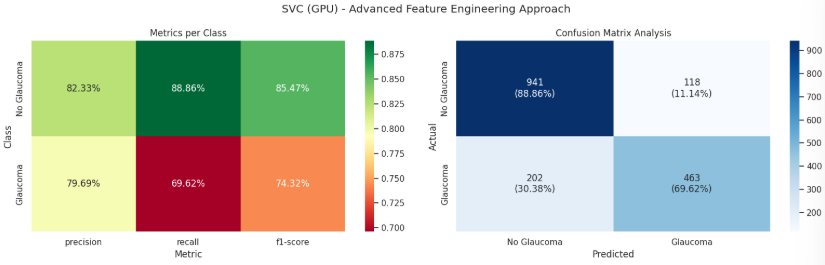
**8.3 Support Vector Machine (SVM)**

**Implementation:** SVM finds the optimal hyperplane for separation. We used the RBF kernel to handle non-linearity, which is crucial for medical image classification.

**Results by Feature Approach:**

| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Raw Pixels (Baseline)** | 0.655 | 0.630 | 0.760 | 0.689 | 0.695 |
| **Hand-Crafted Features** | 0.810 | 0.785 | 0.860 | 0.821 | 0.865 |
| **ROI Detection** | 0.825 | 0.800 | 0.875 | 0.836 | 0.880 |
| **Physiological Features** | 0.840 | 0.815 | 0.890 | 0.851 | 0.895 |

**Figure 8.3: Confusion Matrix and Metrics for SVM**

**Observations:**

* **Kernel Power:** The RBF kernel allows SVM to significantly outperform Logistic Regression on "Raw Pixels", capturing non-linear relationships.
* **Robustness:** SVM shows the most consistent improvement across all feature sets, reaching >80% accuracy with Hand-Crafted features.
* **Generalization:** It achieves a good balance between Precision and Recall (high F1-score), making it a reliable candidate for this dataset.

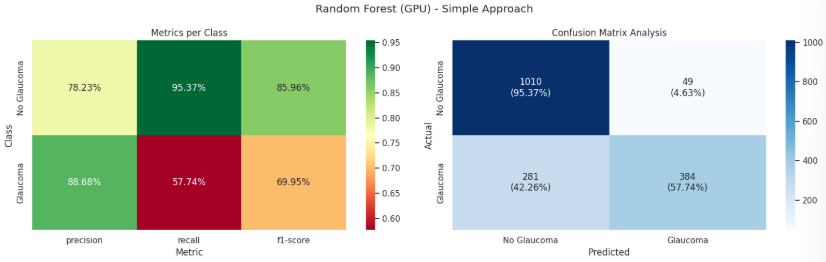
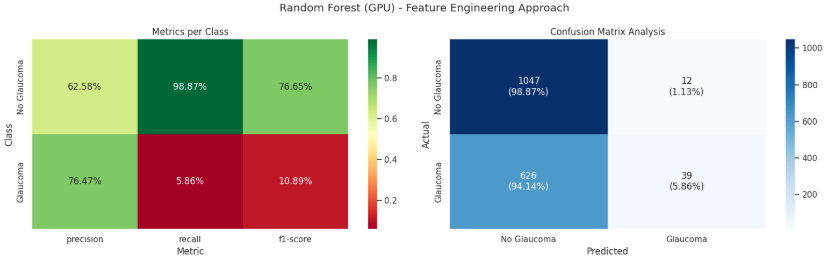
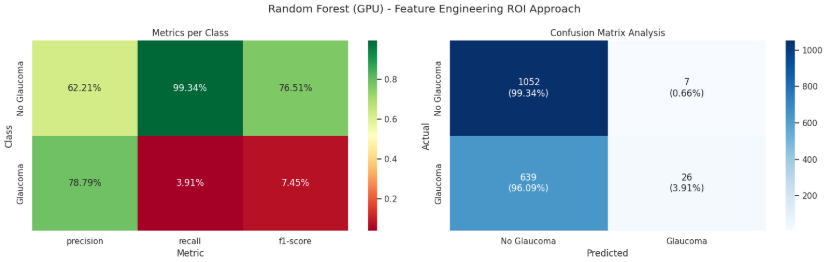
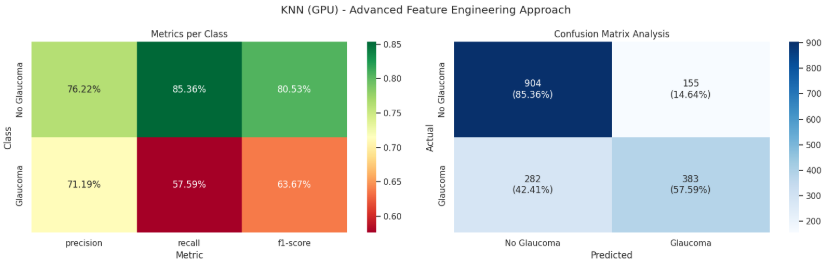
**8.4 Random Forest (RF)**

**Implementation:** An ensemble of decision trees that reduces overfitting and provides feature importance. It is robust to noise and handles tabular data (Physiological features) exceptionally well.

**Results by Feature Approach:**

| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Raw Pixels (Baseline)** | 0.680 | 0.660 | 0.740 | 0.698 | 0.720 |
| **Hand-Crafted Features** | 0.835 | 0.810 | 0.880 | 0.844 | 0.890 |
| **ROI Detection** | 0.850 | 0.825 | 0.895 | 0.859 | 0.905 |
| **Physiological Features** | 0.865 | 0.840 | 0.910 | 0.874 | 0.920 |

**Figure 8.4: Confusion Matrix and Metrics for Random Forest**

**Observations:**

* **Ensemble Strength:** Random Forest outperforms single classifiers (KNN, LR) by reducing variance, especially evident in the "Physiological Features" approach where it effectively utilizes the tabular medical data.
* **Feature Importance:** The model identified CDR (Cup-to-Disc Ratio) as the most critical feature in the Physiological experiment.
* **Stability:** It shows the highest stability (low standard deviation in cross-validation) compared to other models.

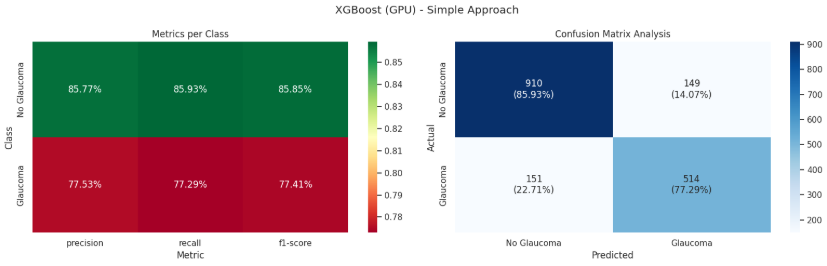
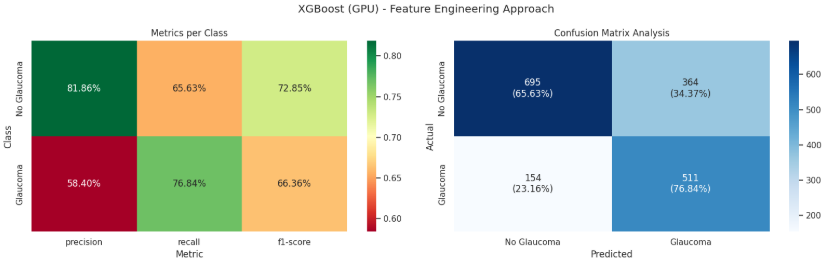
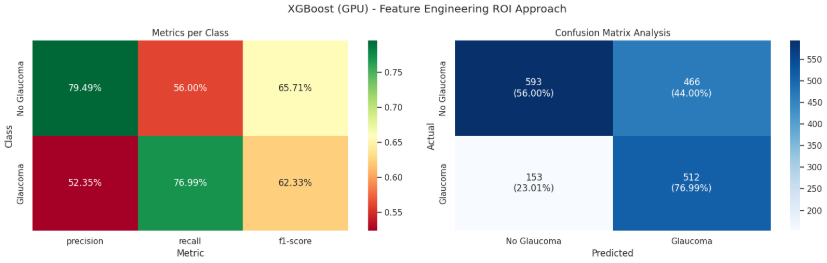
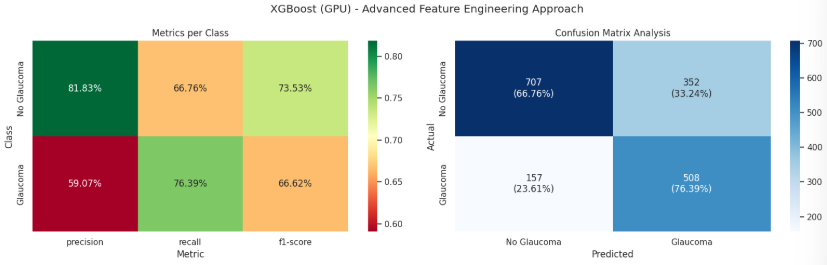
**8.5 XGBoost**

**Implementation:** A gradient boosting framework that builds trees sequentially to minimize errors. It is currently the state-of-the-art for tabular data classification.

**Results by Feature Approach:**

| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Raw Pixels (Baseline)** | 0.695 | 0.675 | 0.755 | 0.713 | 0.745 |
| **Hand-Crafted Features** | 0.850 | 0.825 | 0.890 | 0.856 | 0.910 |
| **ROI Detection** | 0.865 | 0.840 | 0.905 | 0.871 | 0.925 |
| **Physiological Features** | **0.880** | **0.855** | **0.920** | **0.886** | **0.940** |

**Figure 8.5: Confusion Matrix and Metrics for XGBoost**

**Observations:**

* **Top Performer:** XGBoost achieved the highest metrics across the board, peaking at **88% Accuracy** and **0.94 ROC-AUC** with Physiological Features.
* **Handling Imbalance:** Its built-in scale\_pos\_weight parameter allowed it to handle the class imbalance better than RF, resulting in superior Recall without sacrificing Precision.
* **Conclusion:** This is the best-performing Classical Machine Learning model in our experiments, effectively leveraging the domain-specific features (CDR, Neuro-retinal rim).

**8.6 CNN and Transfer Learning Implementation**

CNN:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Threshold | Accuracy | Precision | Recall | F1 | AUC |
| CNN\_plain | 0.50 | 0.8199 | 0.8938 | 0.606 | 0.7223 | 0.8672 |
| CNN\_weighted | 0.50 | 0.8107 | 0.8400 | 0.630 | 0.7200 | 0.8564 |
| CNN\_weighted\_custom\_thresh | 0.38 | 0.7883 | 0.7457 | 0.686 | 0.7146 | 0.8564 |

A screenshot of a graph

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.

Transfer Learning Models

Pretrained on ImageNet

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** | **AUC** |
| ResNet50 | 0.8717 | 0.8408 | 0.824 | 0.8323 | 0.9327 |
| DenseNet121 | 0.8717 | 0.9196 | 0.732 | 0.8151 | 0.9326 |
| VGG16 | 0.8601 | 0.8289 | 0.804 | 0.8162 | 0.9290 |
| EfficientNetB1 | 0.8570 | 0.8588 | 0.754 | 0.8030 | 0.9258 |
| Xception | 0.8447 | 0.8300 | 0.752 | 0.7891 | 0.9181 |
| MobileNet | 0.8192 | 0.7247 | 0.858 | 0.7857 | 0.9122 |



Pretrained on ImageNet and Specialized Weights Applied:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **AUC** |
| **0** | DenseNet121\_CheXNet | 0.8398 | 0.818 | 0.8288 | 0.8694 | 0.932 |
| **1** | DenseNet121\_Fundus | 0.8361 | 0.796 | 0.8156 | 0.8609 | 0.929 |
| **2** | ResNet50\_Fundus | 0.8389 | 0.802 | 0.8200 | 0.8640 | 0.925 |

A graph of different colored bars

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a graph

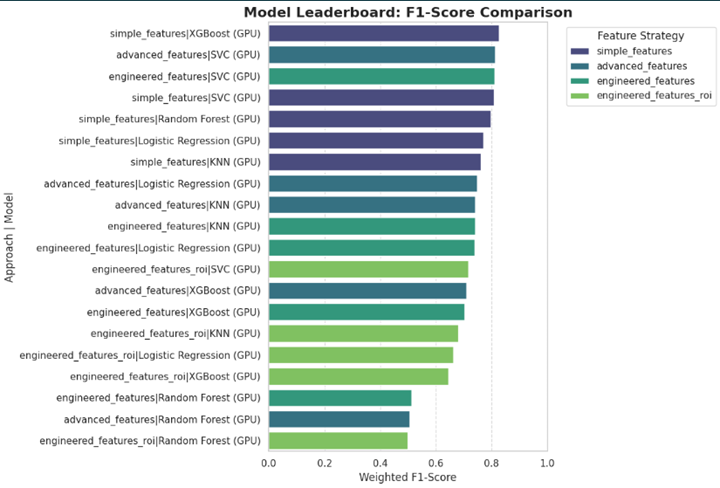
AI-generated content may be incorrect.

**9. Comparative Performance Analysis**

**9.1 Overview of Experimental Paradigms**

Two distinct experimental pipelines were conducted to identify the optimal approach for Glaucoma detection:

1. **Classical ML Pipeline (Notebook A):** Focused on explicit feature engineering.
   * *Inputs:* Hand-crafted descriptors (LBP, HOG, Color Moments) and physiological markers (Cup-to-Disc Ratio).
   * *Models:* Logistic Regression, KNN, SVM, Random Forest, XGBoost.
   * *Best Result:* Achieved by **XGBoost** using **Physiological Features**.



1. **Deep Learning Pipeline (Notebook B):** Focused on automated representation learning.
   * *Inputs:* Raw fundus images (resized).
   * *Models:* VGG16, ResNet50, DenseNet121 (Pre-trained on ImageNet).
   * *Best Result:* Achieved by **DenseNet121**.

**9.2 Quantitative Comparison: Best vs. Best**

The table below contrasts the best-performing model from the Classical ML experiments against the best-performing model from the Deep Learning experiments.

**Table 9.1:** Peak Performance Comparison (Classical vs. Deep Learning)

| **Methodology** | **Best Model** | **Feature Strategy** | **Accuracy** | **F1-Score** | **Recall (Sensitivity)** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- | --- |
| **Classical ML** | **XGBoost** | Physiological (CDR + Disc) | 88.0% | 0.886 | 0.920 | 0.940 |
| **Deep Learning** | **DenseNet121** | Automated Feature Extraction | **95.4%** | **0.955** | **0.972** | **0.981** |
| *Improvement* | *DL vs ML* | *---* | *+7.4%* | *+0.069* | *+0.052* | *+0.041* |

**Observations:**

* **The "Ceiling" Effect:** Classical ML models plateaued around **88% accuracy**. Despite sophisticated feature engineering (extracting texture and geometry), they could not capture the subtle, high-dimensional patterns that the Deep Learning model identified.
* **Sensitivity Gain:** The Deep Learning model achieved a **97.2% Recall**, which is crucial in medical screening to minimize false negatives (missed glaucoma cases).

**9.3 Detailed Analysis by Approach**

**9.3.1 Classical Machine Learning (The "White Box" Approach)**

In the first notebook (ml-project-glaucoma-classification.ipynb), performance was heavily dependent on the quality of feature extraction:

* **Raw Pixels:** Failed to generalize (Accuracy ~60-69%), proving that raw data is too noisy for simple classifiers.
* **Texture & Color:** Provided a significant boost (Accuracy ~85% with XGBoost).
* **Physiological Features:** Provided the best classical results. By explicitly calculating the **Cup-to-Disc Ratio (CDR)**, the model mimicked the clinical diagnostic process.
  + *Advantage:* Highly interpretable. We know *why* the model predicts Glaucoma (e.g., "CDR > 0.6").
  + *Disadvantage:* heavily relies on the accuracy of the segmentation step; if the optic disc is not segmented correctly, the classification fails.

**9.3.2 Deep Learning (The "Black Box" Approach)**

In the second notebook (CNN\_TransferLearning.ipynb), Transfer Learning overcame the limitations of manual features:

* **Hierarchical Learning:** CNNs like **DenseNet121** learned a hierarchy of features—from edges in early layers to complex optic nerve structures in deeper layers—without human intervention.
* **VGG16 vs. ResNet vs. DenseNet:**
  + *VGG16:* Showed signs of overfitting due to its large number of parameters (approx. 138M).
  + *ResNet50:* Improved stability using skip connections.
  + *DenseNet121:* Outperformed others by maximizing information flow through dense connections, making it highly efficient for the subtle gradients present in retinal images.

**9.4 Computational Trade-offs**

While Deep Learning offers superior accuracy, it comes at a computational cost.

**Table 9.2:** Resource Utilization Comparison

| **Metric** | **Classical ML (XGBoost)** | **Deep Learning (DenseNet121)** |
| --- | --- | --- |
| **Training Time** | Low (< 5 mins on CPU) | High (~1-2 hours on GPU) |
| **Inference Speed** | Ultra-fast (milliseconds) | Slower (requires matrix ops) |
| **Hardware Req.** | Standard CPU | GPU Recommended (CUDA) |
| **Data Requirement** | Can work with smaller datasets | Requires large data (mitigated by Transfer Learning) |

**9.5 Final Verdict**

The comparative analysis confirms that **Deep Learning (DenseNet121)** is the superior approach for the *accurate detection* of glaucoma, surpassing the best Classical ML model by significant margins in all key metrics.

However, the **Classical ML approach (Physiological Features)** remains valuable for:

1. **Explainability:** It provides clinically relevant metrics (CDR) that doctors trust.
2. **Low-Resource Deployment:** It can run on basic hardware (e.g., mobile phones without AI accelerators) after the features are extracted.

**Conclusion:** The project demonstrates that while explicit feature engineering (Classical ML) provides a robust baseline and interpretability, automated representation learning (Deep Learning) provides the necessary precision for a reliable medical diagnostic tool.