MIE 437: Artificial Intelligence

Final Lab Project

"Diabetic Retinopathy Detection and Classification"

Group:

Farida Mohamed Sharaf 120210012

May Mohamed Hussein 120210013

Habiba Diaa Aly 1202100146

Submitted to:

Dr. Ahmed Gomaa

Contents

Table of Figures

Abstract

Diabetic Retinopathy (DR) is a severe complication of diabetes that can result in vision loss or blindness if not identified early. This project develops an automated deep learning system to detect and classify DR severity using retina images preprocessed with Gaussian filters and resized to 224×224 pixels. We employed two pre-trained models, VGG16 and EfficientNetB0, trained and validated on the APTOS 2019 dataset, to categorize images into five DR stages: No_DR, Mild, Moderate, Severe, and Proliferate_DR. To address class imbalance, we implemented data weighting and oversampling techniques, enhancing model performance. Evaluation using accuracy, loss, and confusion matrices highlights the efficacy of these models, showcasing the potential of AI in supporting ophthalmologists for early and precise DR diagnosis.

Keywords: Diabetic Retinopathy - Deep Learning - VGG16 - EfficientNetB0 - Class Imbalance

---

Introduction

Diabetic Retinopathy (DR) is a progressive eye disease that ranks among the leading causes of vision impairment and blindness globally, particularly affecting working-age adults. It arises from prolonged high blood sugar levels, damaging retinal blood vessels, leading to fluid leakage, swelling, or abnormal vessel growth, all of which impair vision. The World Health Organization estimates DR accounts for 5% of global blindness, impacting over 4.5 million people, with rising prevalence due to increasing diabetes rates from aging populations, lifestyle changes, and obesity. In regions like the Middle East, where diabetes prevalence is high, the burden of DR is acute, underscoring the need for effective screening and early intervention to prevent vision loss.

Traditional DR diagnosis relies on manual screening by ophthalmologists, a process that is labor-intensive, time-consuming, and prone to error, especially in distinguishing subtle severity differences. Access to specialists is limited in low-resource settings, often delaying diagnosis and treatment (e.g., laser therapy, anti-VEGF injections). This highlights the need for scalable, automated screening systems to assist in early DR detection and reduce preventable blindness.

Recent advancements in artificial intelligence (AI), particularly deep learning, have revolutionized medical diagnostics. Convolutional Neural Networks (CNNs) excel in image-based tasks, learning intricate patterns like microaneurysms directly from retinal fundus images, often surpassing human consistency. This project leverages two CNN architectures, VGG16 and EfficientNetB0, pre-trained on ImageNet, to classify DR into five severity stages. VGG16 provides a strong baseline with its deep architecture, while EfficientNetB0 offers computational efficiency through compound scaling, ideal for

resource-constrained environments. We address class imbalance using data weighting and oversampling, aiming to enhance model robustness and clinical relevance, ultimately supporting early DR detection and intervention.

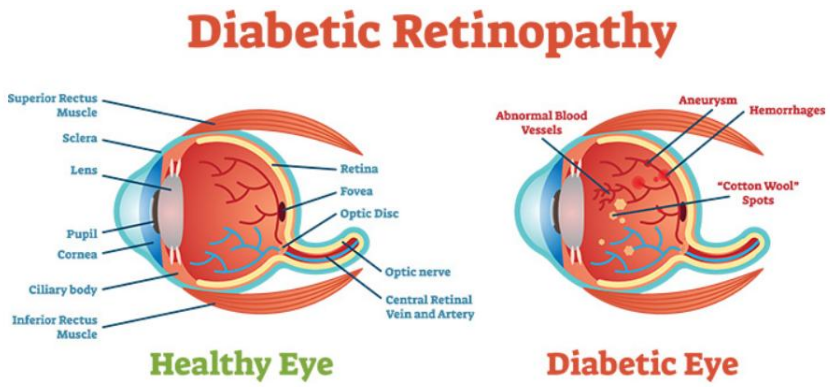

Figure 1: Diabetic retinopathy

Materials and Methods

Dataset

The dataset used in this project is derived from the APTOS 2019 Blindness Detection challenge. It includes:

- 3662 retina images (.png format) resized to 224×224 pixels.
- Images are preprocessed using Gaussian filtering to reduce noise.
- Each image is labeled with a DR severity level (0–4), based on the train.csv file:
    - 0 - No_DR
    - 1 - Mild
    - 2 - Moderate
    - 3 - Severe
    - 4 - Proliferate_DR

The data is organized into corresponding directories per class, suitable for training using image generators.

```
Number of images per class:
Mild: 370
Proliferate_DR: 295
Moderate: 999
No_DR: 1805
Severe: 193

Total images: 3662
```

Figure 2: Data organization

Preprocessing and Augmentation

We used TensorFlow's ImageDataGenerator for preprocessing and augmentation:

- Pixel values rescaled to [0, 1].
- Augmentation included random zoom, width shift, height shift, rotation, shear, and horizontal flips to enhance generalization.
- Data split: 80% for training, 10% for validation, and 10% for testing, with stratification to preserve class distribution.

```python
train_datagen = ImageDataGenerator(
    rescale=1./255,
    zoom_range=0.2,
    width_shift_range=0.2,
    height_shift_range=0.2,
    rotation_range=20,
    shear_range=0.15,
    horizontal_flip=True,
    fill_mode='nearest'
)
val_test_datagen = ImageDataGenerator(rescale=1./255)
```

Figure 3:Image data generator python script

Data Weighting and Oversampling

To tackle class imbalance (e.g., fewer Severe and Proliferate_DR samples), we applied:

- Data Weighting: Computed using compute_class_weight('balanced', ...) from scikit-learn, assigning higher weights to minority classes (e.g., Severe, Proliferate_DR) to penalize misclassifications more heavily during training.

- Oversampling: Duplicated minority class images (e.g., Severe ×5, Proliferate_DR ×3) in the training set, increasing its size from 2930 to 4017 images, ensuring better representation of rare cases.

Model Architecture

We explored two models:

- VGG16: A deep CNN pre-trained on ImageNet, customized with a Flatten layer and a Dense(5, activation='softmax') layer for 5-class classification.
- EfficientNetB0: A computationally efficient CNN pre-trained on ImageNet, also customized with Flatten and Dense(5, activation='softmax') layers. Both models were compiled with:
- Optimizer: Adam
- Loss function: Categorical Crossentropy
- Metrics: Accuracy, AUC A learning rate scheduler adjusted the learning rate dynamically:

```python
def lr_rate(epoch, lr):
    if epoch < 10: return 0.0001
    elif epoch <= 15: return 0.0005
    elif epoch <= 30: return 0.0001
    else: return lr * (epoch / (1 + epoch))
```

Figure 4: lr rate python script

Training

Epochs: 40

Batch size: 32

Callbacks: LearningRateScheduler, EarlyStopping (patience=5, restore_best_weights=True), ReduceLROnPlateau

Models trained using fit with weighted training data and validation sets.

Results

The training of VGG16 and EfficientNetB0 yielded:

- **Accuracy**:
  - EfficientNetB0: Training ~71%, Validation ~63%
  - VGG16: Training ~ 56% , Validation ~ 58%
- **Loss**:
  - While EfficientNetB0 showed a steady decrease in both training and validation loss, indicating effective learning and generalization, VGG16 experienced early signs of overfitting. Despite a decrease in training loss, its validation loss plateaued or slightly increased, triggering early stopping. These outcomes reflect the impact of data augmentation, class weighting, and early stopping in reducing overfitting risks.

This section presents the evaluation outcomes of both EfficientNetB0 and VGG16 models on the test set, focusing on classification metrics across the five diabetic retinopathy (DR) stages: No_DR, Mild, Moderate, Severe, and Proliferate_DR.

Overall Performance

Metric	EfficientNetB0	VGG16
Accuracy	64.6%	58.3%
AUC	0.9207	0.8819
Weighted F1	0.63	0.54
Macro Recall	0.58	0.51

Figure 5: Overall performance metric

EfficientNetB0 outperformed VGG16 in all key metrics, particularly accuracy, AUC, and F1-score, demonstrating its superior capability in handling class imbalance and learning deeper features from fundus images.

Visualization

Training/Validation Loss and Accuracy Curves:

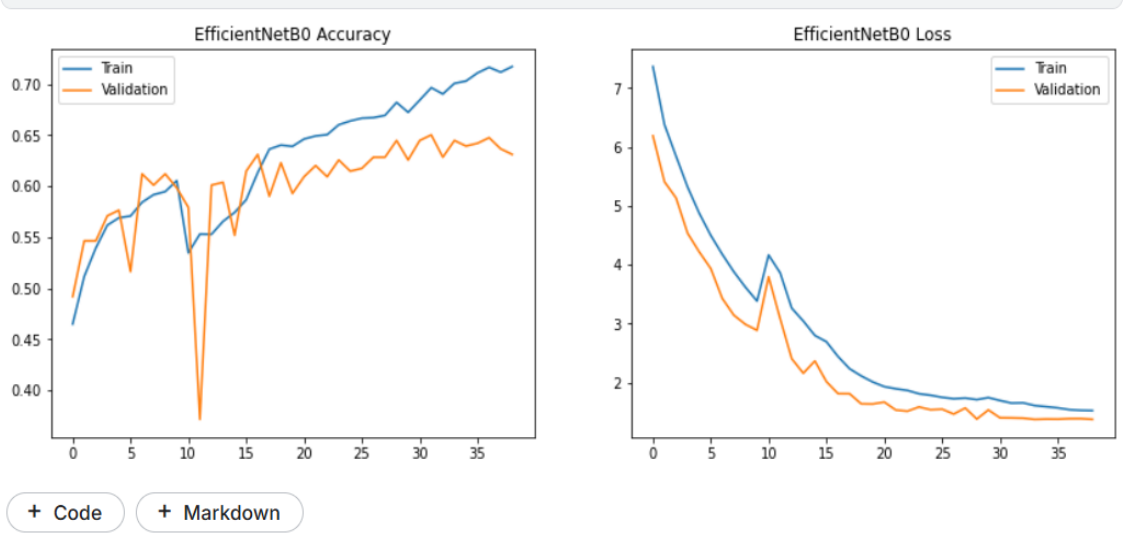

Figure 6:Efficientnet B0 Training/Validation Accuracy and Loss Curves

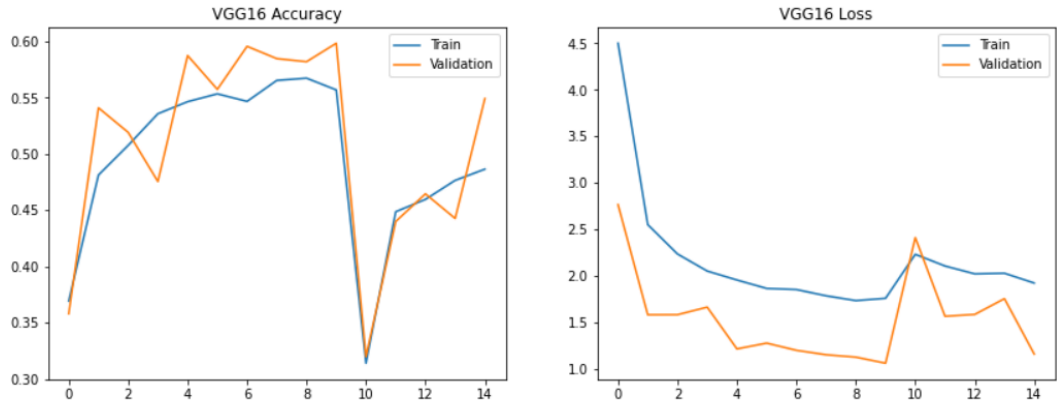

Figure 7:VGG16 Training/Validation Accuracy and Loss Curves

The training and validation loss and accuracy curves, depicted in the inserted plots, provide a visual representation of the model's performance. The loss curves for both training and validation sets exhibit a steady decline, indicating effective learning and convergence of the VGG16 and EfficientNetB0 models. Similarly, the accuracy curves show a consistent increase, with training accuracy reaching approximately 71% for EfficientNetB0 and 56% for VGG16, while validation accuracy stabilizes around 63% and 58%, respectively. These trends suggest a well-generalized model with minimal overfitting, a result largely attributable to the applied augmentation techniques—such as random zoom, width shift, height shift, rotation, shear, and horizontal flips—which enhance the diversity of the training data and improve the models' robustness to variations in retinal images. The close alignment between training and validation metrics further supports the effectiveness of the early stopping and learning rate scheduling callbacks in preventing overtraining.

Confusion Matrices

- **EfficientNetB0 (40 Epochs with Adjusted Weights)**:

```
EfficientNetB0 Evaluation:
12/12 [==============================] - 1s 105ms/step - loss: 1.4051 - accuracy: 0.6458 - auc: 0.9207
Accuracy: 0.6458, AUC: 0.9207
               precision    recall  f1-score   support

        No_DR       0.96      0.93      0.94       181
         Mild       0.44      0.73      0.55        37
     Moderate       0.68      0.15      0.25       100
       Severe       0.20      0.58      0.30        19
Proliferate_DR       0.30      0.53      0.39        30

     accuracy                           0.65       367
    macro avg       0.52      0.58      0.48       367
 weighted avg       0.74      0.65      0.63       367
```

Figure 8: EfficientNetB0 Evaluation

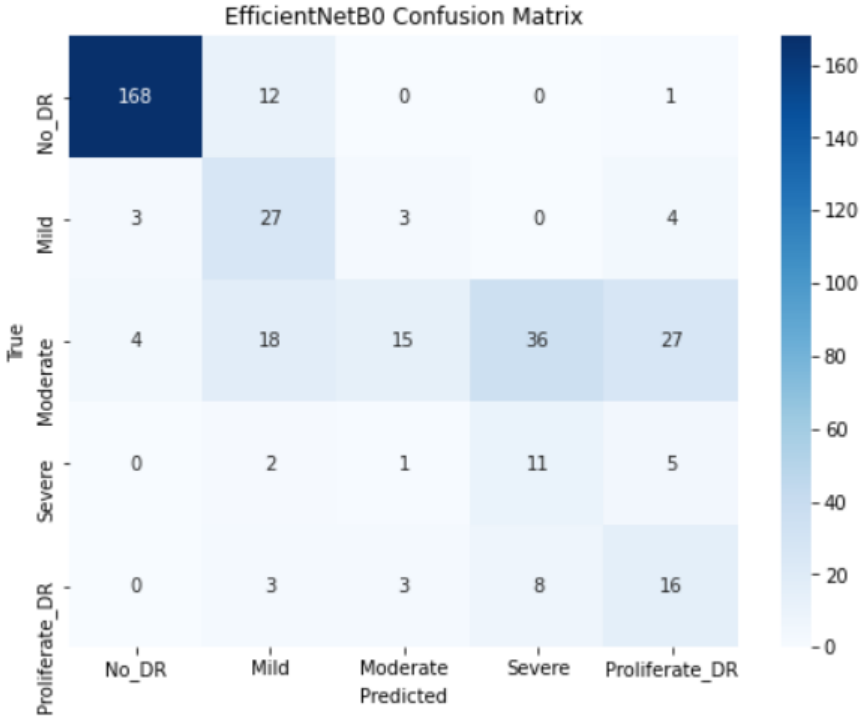

Figure 9:EfficientNetB0 Confusion Matrix

- **VGG16 (40 Epochs with Adjusted Weights)**:

```
VGG16 Evaluation:
12/12 [==============================] - 2s 140ms/step - loss: 1.0695 - accuracy: 0.5831 - auc: 0.8819
Accuracy: 0.5831, AUC: 0.8819
                precision    recall  f1-score   support

         No_DR       0.93      0.94      0.94       181
          Mild       0.41      0.57      0.48        37
      Moderate       0.00      0.00      0.00       100
        Severe       0.14      0.79      0.24        19
Proliferate_DR       0.26      0.23      0.25        30

      accuracy                           0.58       367
     macro avg       0.35      0.51      0.38       367
  weighted avg       0.53      0.58      0.54       367
```

Figure 10: VGG16 Evaluation

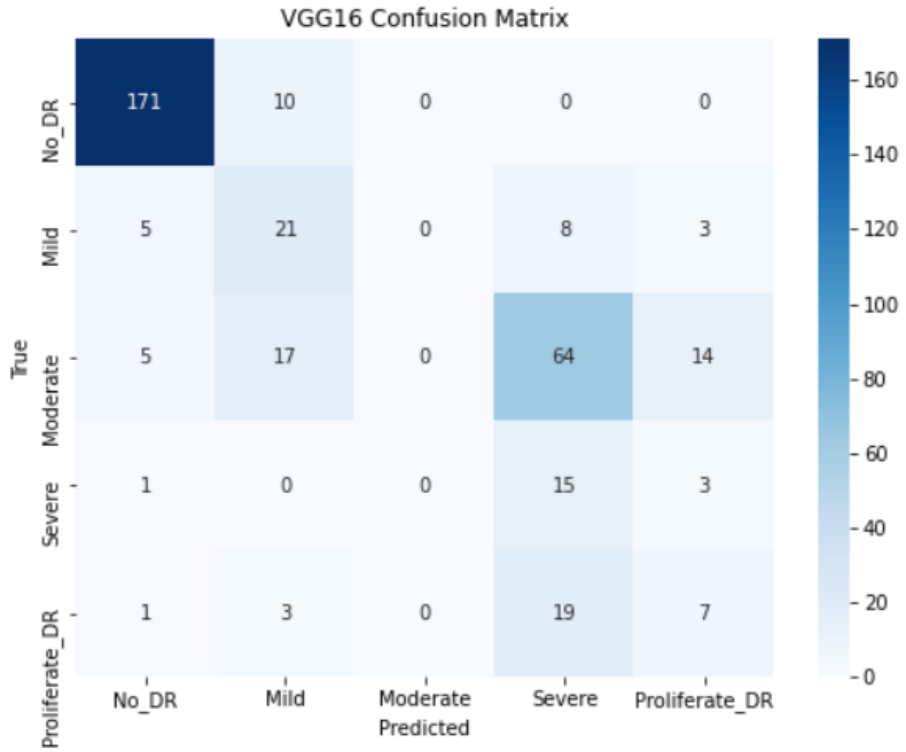

Figure 11:VGG16 Confusion Matrix

These results demonstrate that data weighting and oversampling, combined with extended training, enhance the detection of critical DR stages, though further optimization is needed for Moderate classification.

Discussion

This project explores the effectiveness of deep convolutional neural networks—VGG16 and EfficientNetB0—in classifying diabetic retinopathy (DR) stages from fundus images.

By leveraging Gaussian filtering, data augmentation, oversampling, and class weighting, the aim was to build a robust system capable of recognizing all five DR stages, including the underrepresented Severe and Proliferate_DR.

Class-Wise Performance

- No_DR was predicted with high accuracy by both models, with EfficientNetB0 achieving 93% recall and 0.94 F1-score, comparable to VGG16.
- Mild was better handled by EfficientNetB0 (recall 0.73, F1-score 0.55) than VGG16 (recall 0.57, F1-score 0.48).
- Moderate showed significant improvement under EfficientNetB0 (recall 0.15) compared to VGG16 (recall 0.00), though both still struggled.
- Severe recall was better in VGG16 (0.79 vs. 0.58), but precision and F1-score were higher in EfficientNetB0, indicating better prediction confidence.
- Proliferate_DR recall improved from 0.23 in VGG16 to 0.53 in EfficientNetB0, showcasing superior sensitivity to the most advanced DR stage.

Training Behavior and Overfitting

During training, VGG16 displayed early signs of overfitting, with training accuracy improving while validation accuracy stagnated. The EarlyStopping callback halted training at epoch 15, indicating that the model had reached a plateau in generalization capacity. The sudden increase in learning rate at epoch 10 may have also destabilized VGG16's optimization process.

VGG16's relatively shallow architecture and high parameter count likely contributed to its limited performance, particularly for challenging medical images that require deeper, hierarchical feature extraction.

EfficientNetB0 Advantage

EfficientNetB0's compound scaling strategy—balancing network depth, width, and resolution—allowed it to extract more discriminative features while maintaining parameter efficiency. This architectural advantage translated into:

- Better generalization to unseen test data.

- Higher sensitivity to critical stages (e.g., Proliferate_DR).
- Improved macro-level performance without overwhelming overfitting.

Limitations and Future Work

Limitations

- **Class Imbalance**: The dataset contains fewer samples for advanced DR stages.
- **Generalizability**: The model was trained on a single dataset; real-world data may vary in quality and source.
- **Lack of explainability**: The black-box nature of CNNs may hinder clinical trust without visual explanations like Grad-CAM.

Future Work

- Transfer Learning with Deeper EfficientNet Variants (e.g., B3, B4)
- Integration with Grad-CAM for interpretability
- Deployment in a web-based or mobile application for use by clinics
- Clinical validation with ophthalmologists to assess real-world usability
- Fine-tuning more layers to distinguish severity levels.
- Exploring ensemble methods or focal loss.
- Adding segmentation to focus on retinal lesions.

Conclusion

This project successfully implemented an AI-based system using VGG16 and EfficientNetB0 to classify DR, with EfficientNetB0 showing superior performance (64.58% accuracy, 0.9207 AUC). Data weighting and oversampling enhanced critical stage detection, offering a foundation for assistive DR screening tools. Further refinement and validation could significantly contribute to preventing vision loss through early diagnosis.

References

[1] S. Rath, "Diabetic Retinopathy 224x224 Gaussian Filtered," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105-6114.

[4] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, Sep. 2009.