Deep Learning-Based Multi-Class Classification of Diabetic Retinopathy Using Fundus Images

Jeeva Jose C
Batch 11

1. Introduction

Diabetic retinopathy (DR) is a serious complication of diabetes that can cause irreversible vision loss if not diagnosed and treated early. Traditional manual screening methods are often time-consuming, subjective, and require specialized expertise. Therefore, automated deep learning-based systems have great potential to assist healthcare professionals in the early detection and classification of DR.

In this study, a deep learning approach using Convolutional Neural Networks (CNNs) is employed to classify fundus images into multiple stages of diabetic retinopathy. The model is trained, validated, and tested using a public dataset, and its performance is evaluated using multiple metrics to ensure reliability.

2. Data Preparation and Label Assignment

Before training the model, the images and labels were prepared through several steps:

- **Image Loading and Preprocessing:** Each image was read, converted into an array, resized to 256x256 pixels to ensure uniform size, and normalized to improve convergence during training.
- **Label Assignment:** The images were organized in folders where specific characters in the file path indicated the class label. Based on the folder names:
  - ‘1’ → Class 0: No DR signs
  - ‘4’ → Class 1: Severe NPDR
  - ‘5’ → Class 2: Very Severe NPDR
  - ’7’ → Class 3: Advanced PDR
- **One-Hot Encoding:** The class labels were converted into categorical format suitable for multi-class classification.
- **Shuffling and Splitting:** The entire dataset was randomly shuffled to avoid any order bias, and then split into 80% training, 20% temporary set (which was further split into validation and test sets) using stratified sampling to preserve class distribution.

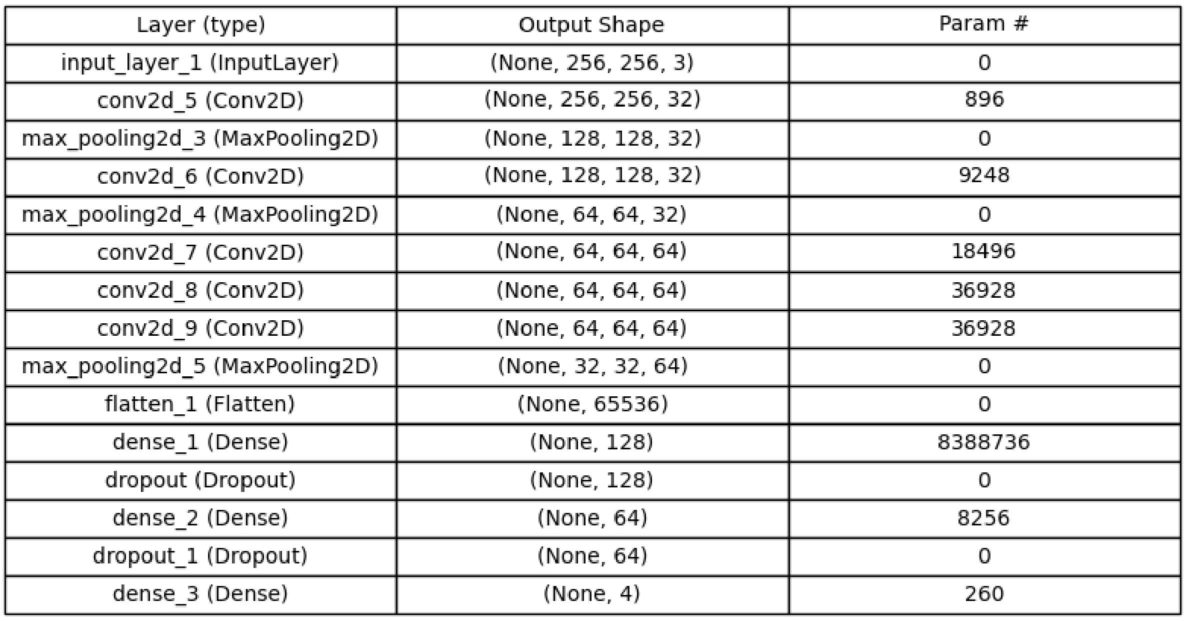
3. Model Architecture

The classification model is based on a Convolutional Neural Network (CNN), inspired by AlexNet, but adapted to this task. It processes the input fundus images (256x256x3) and automatically learns important features useful for classification.

The architecture includes:

- Two convolutional layers with 32 and 64 filters, each followed by max pooling to reduce spatial dimensions while retaining key features like edges, textures, and lesions.
- A flattening layer to convert extracted features into a one-dimensional vector.
- Two fully connected (dense) layers with 128 and 64 neurons to learn high-level patterns.
- Dropout layers (rates of 0.3 and 0.2) between dense layers to prevent overfitting.
- A final softmax output layer with 4 neurons for class prediction.

This simple yet effective architecture balances model complexity and generalization ability, making it suitable for medical image analysis with limited data.



Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 256, 256, 3)	0
conv2d_5 (Conv2D)	(None, 256, 256, 32)	896
max_pooling2d_3 (MaxPooling2D)	(None, 128, 128, 32)	0
conv2d_6 (Conv2D)	(None, 128, 128, 32)	9248
max_pooling2d_4 (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_7 (Conv2D)	(None, 64, 64, 64)	18496
conv2d_8 (Conv2D)	(None, 64, 64, 64)	36928
conv2d_9 (Conv2D)	(None, 64, 64, 64)	36928
max_pooling2d_5 (MaxPooling2D)	(None, 32, 32, 64)	0
flatten_1 (Flatten)	(None, 65536)	0
dense_1 (Dense)	(None, 128)	8388736
dropout (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
dropout_1 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 4)	260

Figure 1: Model Summary

Total parameters: 8,499,748 (32.42 MB)
Trainable parameters: 8,499,748 (32.42 MB)
Non-trainable parameters: 0 (0.00 B)

4. Model Compilation, Data Augmentation, and Training

The model was compiled using the Adam optimizer, which adaptively adjusts the learning rate for efficient convergence. Categorical cross-entropy was used as the loss function, suitable for multi-class classification.

To improve generalization, data augmentation was applied during training. The following transformations were used:

- Random rotations (±20°)
- Horizontal and vertical shifts (up to ±10%)
- Random zoom (up to ±10%)
- Horizontal flipping

These augmentations simulate real-world variations in fundus images and help the model become robust to different orientations and scales.

Custom Evaluation Metrics

In addition to accuracy, precision, and recall, two additional metrics were implemented to better evaluate the model's performance:

- **Dice Coefficient:** Measures the overlap between predicted and true labels, commonly used in medical imaging.

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- **Intersection over Union (IoU):** Calculates the ratio of correctly predicted overlap to the total area covered by the prediction and ground truth.

$$IoU = \frac{TP}{TP + FP + FN}$$

The model was trained for 100 epochs with a batch size of 32. The best model was selected based on the highest validation accuracy using model checkpointing.

5. Model Evaluation and Results

The model's performance was evaluated on training, validation, and test sets using accuracy, loss, precision, recall, Dice coefficient, and IoU. The best validation accuracy was achieved at epoch 97. Table 1 summarizes the complete performance:

Table 1: Model Performance on Training, Validation, and Test Sets

Dataset	Accuracy	Loss	Dice	IoU	Precision	Recall
Training	82.26%	0.4938	0.7419	0.5920	87.56%	78.21%
Validation	77.59%	0.9650	0.7035	0.5630	81.13%	74.14%
Test	74.58%	0.6332	0.7032	0.5435	77.78%	71.19%

The training accuracy indicates effective learning, while the validation and test results demonstrate good generalization. Although the validation loss is higher than training loss, the consistent precision, recall, Dice, and IoU values indicate balanced and reliable performance across different datasets.

Loss Curve Analysis

The training and validation loss curves were plotted to observe the learning behavior of the model. As seen in Figure 2, both training and validation losses fluctuated during training. This indicates that while the model was able to learn from the data, it also faced some difficulty in maintaining stable improvement, likely due to the limited dataset size and inherent variability in medical images.

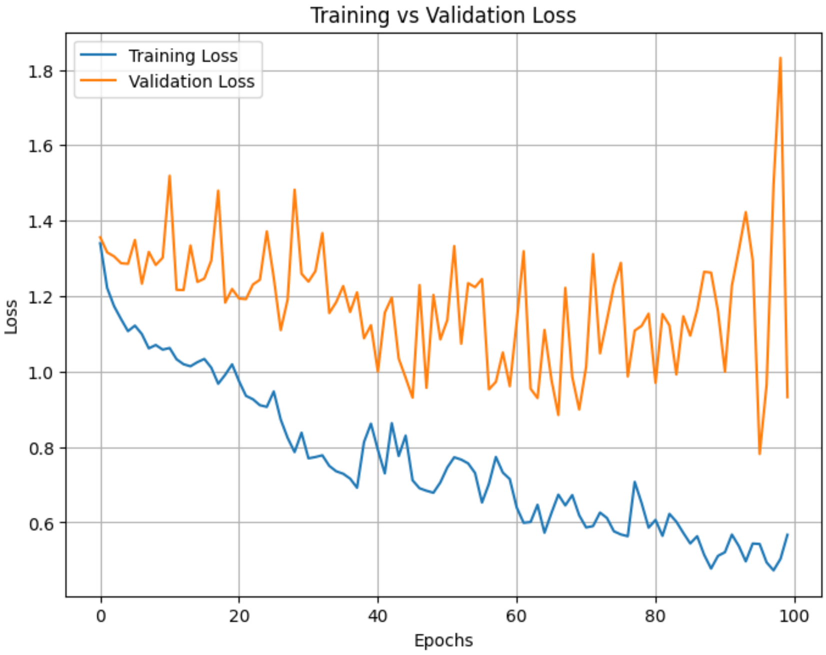


Figure 2: Training vs Validation Loss

6. Conclusion

In this study, a deep learning-based approach was successfully implemented for the multi-class classification of diabetic retinopathy using fundus images. The model demonstrated good performance across training, validation, and test sets, achieving a maximum validation accuracy of 77.59% and test accuracy of 74.58%.

The use of CNNs allowed the model to automatically extract important features relevant to different stages of DR. The addition of Dice and IoU metrics provided more insight into the model's ability to correctly capture disease-specific regions.

Although the results are promising, further improvements can be achieved by training on larger and more diverse datasets, experimenting with more complex architectures, and incorporating explainable AI techniques for clinical deployment.