Diabetic Retinopathy Detection

Flavio Caroli

***Abstract*—Diabetic retinopathy (DR) is a global health concern and a major cause of blindness. Early detection is crucial but hindered by the time-intensive process of manual fundus image analysis. Machine learning provides an opportunity to automate and enhance the efficiency of DR screening.**

I. Introduction

Diabetic retinopathy (DR) is a leading cause of vision impairment, necessitating early detection for effective intervention. Manual screening is labor-intensive and delays diagnosis. In this study, I leverage the APTOS 2019 dataset of 3,662 retinal images, addressing challenges such as class imbalance and inter-class similarity. By employing efficient neural network architectures and preprocessing techniques, I aim to develop a robust framework for automated DR detection.

II. Dataset Exploration

Understanding the dataset's structure and addressing challenges like class imbalance were key to this study.

A. Data Structure and Integrity

The dataset included 3,296 training and validation samples, with each sample containing:

- **id_code**: Unique image identifier.
- **diagnosis**: DR severity level (0: no DR, 4: proliferative DR).

Thorough checks confirmed no overlap between training and validation sets. **5-fold stratified cross-validation** was implemented to preserve class proportions and ensure robust model evaluation.

B. Class Distribution

Analysis revealed significant class imbalance:

- **Class 0 (no DR)**: $\approx 50\%$ of the dataset.
- **Class 2 (moderate DR)**: $\approx 28\%$.
- **Classes 1, 3, 4**: Severely underrepresented.

This imbalance necessitated techniques such as class weighting to address bias and improve performance on minority classes.

III. Preprocessing and Image Preparation

Preprocessing was essential to emphasize diagnostic features and prepare the dataset for effective training and evaluation.

A. Reflection on Dataset Challenges

The APTOS 2019 dataset, presents inherent biases due to meta features like aspect ratio, height, width, and pixel count [1]. These meta features can disproportionately influence predictions, allowing models to exploit spurious correlations instead of learning diagnosis-relevant patterns. Key challenges include:

- **Bias from Meta Features**: Meta features, such as square aspect ratios (1.0), correlate with certain classes (e.g., class 0), enabling the model to predict labels based on image structure rather than diagnostic content.
- **Overfitting Risks**: Without proper preprocessing, models can achieve over 90% accuracy by leveraging these biases, failing to generalize to unseen test data.

To mitigate these issues is critical:

- **Preprocessing techniques**, such as cropping and blurring, help focus models on retinal features rather than meta patterns.
- **Augmentation** introduces diversity in image dimensions and distributions, reducing over-reliance on spurious features.

B. Preprocessing Pipeline

The preprocessing pipeline, adapted from Ben Graham's method [2], comprised:

- **Resizing**: Standardized images to 256×256 dimensions, ensuring uniform input size for model training.
- **Cropping**: Removed uninformative borders to focus on relevant retinal regions.
- **Local Contrast Enhancement**: Applied a local contrast enhancement effect using Gaussian filtering, improving the visibility of retinal structures and fine details for diagnostic tasks.

IV. Training Framework

A modular and scalable training framework was implemented to ensure efficient model training, evaluation, and reproducibility.

A. Trainer Class

The `Trainer` class was developed to manage the training and validation process. Key functionalities included:

- **Data Management**: Initialization of training and validation data loaders with appropriate augmentations.
- **Optimization**: Use of the `Adam` optimizer with dynamic learning rate schedulers, such as `CosineAnnealing` and `ReduceLROnPlateau`.

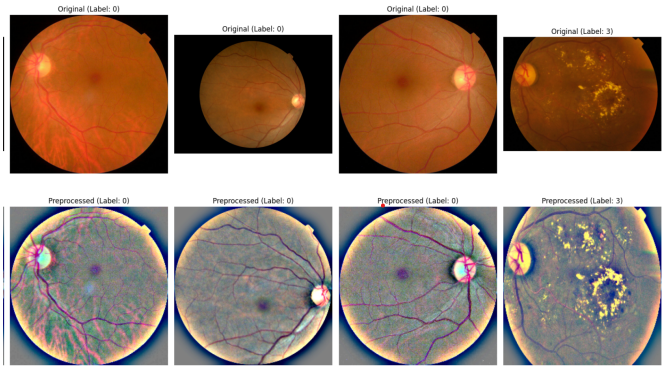

Fig. 1. Figure illustrates the impact of preprocessing, showing clearer and more uniform retinal images.

- **Loss Functions**: Support for `CrossEntropyLoss` and `FocalLoss` to handle class imbalance.
- **Metrics Tracking**: Computation of key metrics, including:
  - **Balanced Accuracy (BACC)**: Measures performance sensitivity to class imbalance.
  - **Quadratic Weighted Kappa (QWK)**: Evaluates agreement in ordinal classification.
  - **Accuracy**: Reflects overall correctness.

The best-performing model for each fold was saved based on validation BACC, ensuring reproducibility and reliability.

B. Configuration Class

The `Config` class centralized hyperparameter management to maintain consistency across experiments. Configurations included:

- **Model and Dataset Parameters**: Batch size, learning rate, input size, and EfficientNet variants (`b0`, `b4`).
- **Training Settings**: Number of epochs, optimizer parameters, and options for unfreezing backbone blocks.
- **Validation Strategy**: Use of 5-fold cross-validation to ensure robust evaluation.

C. Cross-Validation Workflow

A `run_cross_validation` function automated training and validation across folds:

1) **Initialization**: Configuration of fold-specific data splits, class weights, and model hyperparameters.
2) **Training**: Execution of the training loop with data augmentation and real-time metric logging.
3) **Evaluation**: Computation of fold-specific metrics (BACC, QWK, Accuracy).
4) **Model Saving**: Storage of the best-performing model for each fold.

V. Baseline Model Results

The baseline model, an EfficientNet-B0 architecture, was trained using preprocessed images resized to 256×256 and basic augmentations (e.g., rotation, flipping). The evaluation metrics demonstrated the following performance across 5 folds:

- **Mean Balanced Accuracy (BACC)**: $56.50\% \pm 1.82\%$
- **Mean Kappa**: $83.44\% \pm 1.21\%$
- **Mean Accuracy**: $78.70\% \pm 0.60\%$

Despite achieving high overall accuracy, misclassification rates were significant for underrepresented classes, as shown in Figure 2. Key observations include:

- **Class 0**: Dominated predictions, with only 1.55% misclassification.
- **Classes 1 and 2**: Misclassification rates of 47.06% and 21.98%, respectively, due to subtle inter-class similarities (Figure 6).
- **Classes 3 and 4**: Misclassification exceeded 70%, highlighting challenges with minority classes.

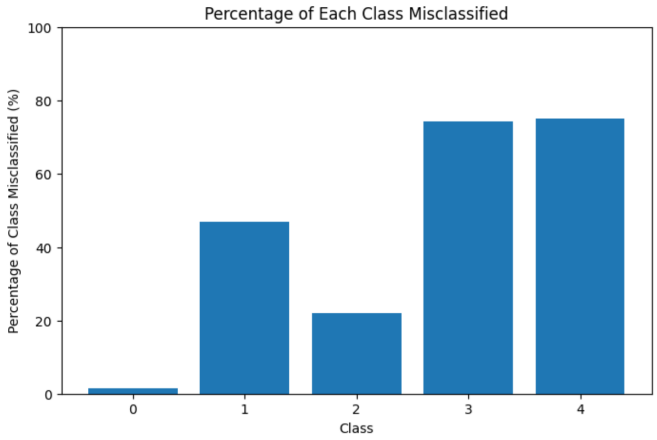

Fig. 2. Percentage of misclassified samples per class. Minority classes (3 and 4) exhibit the highest error rates.

A. Calibration and Reliability

The Expected Calibration Error (ECE) was calculated to assess the reliability of predicted probabilities, yielding a low value of 0.029, indicating well-calibrated outputs suitable for real-world deployment.

B. Visualization and Insights

Figure 4 illustrates examples of correct and incorrect predictions. Analysis of validation metrics and misclassification patterns revealed that the model relied heavily on dominant class predictions, limiting its ability to generalize to minority classes. Improving differentiation between visually similar classes remains a key challenge.

VI. Baseline + Class Weight

Incorporating class weights into the baseline model improved performance for minority classes while maintaining calibration reliability.

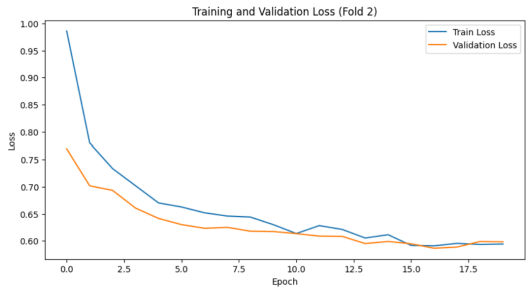

Fig. 3. Training and Validation Loss for the Baseline Model.

A. Key Observations

- **Balanced Accuracy**: Improved to $63.24\% \pm 1.57\%$, indicating better performance across all classes.
- **Class-Specific Gains**:
  - **Class 1**: Misclassification reduced significantly.
  - **Class 3 and 4**: Substantial improvement. in handling this minority class.
  - **Class 2**: The misclassification increased significantly.
- **Class 0**: Slight decrease in accuracy, but maintained strong performance.
- **Calibration**: ECE remained low at 0.0303, reflecting reliable probability outputs.

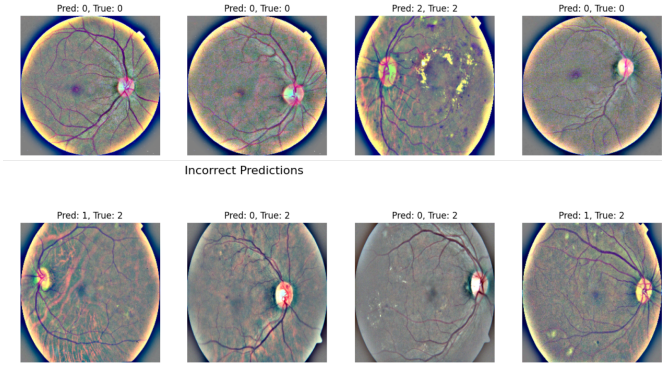

Fig. 4. Examples of correct and incorrect predictions in the Baseline Model. Correct classifications are primarily from dominant classes.

B. Trade-offs and Evaluation

Class weights improved minority class performance at the cost of slight decreases in overall accuracy, which dropped to $73.67\% \pm 1.77\%$. A bootstrap hypothesis test confirmed the statistical significance of these gains:

- **p-value**: < 0.0001.
- **95% Confidence Interval**: [0.0461, 0.0891].

VII. Best Model: Class Weight + Unfreeze 2 Blocks

Unfreezing the last two blocks of the EfficientNet backbone further improved the model's performance by allowing fine-tuning of deeper layers.

A. Key Observations

- **Balanced Accuracy**: Improved to $63.93\% \pm 1.80\%$, surpassing the previous model.
- **Class-Specific Performance**:
  - **Class 1**: Significant gains with reduced misclassification and higher F1-scores.
  - **Class 4**: Marked improvement in F1-score, though challenges remain.
  - **Class 3**: Slight performance decline.
  - **Class 0**: Maintained exceptional accuracy.
- **Calibration**: ECE of 0.0317, ensuring reliable probability outputs.

B. Statistical Significance and Trade-offs

The improvements in balanced accuracy were statistically significant. While macro-averaged F1-score and recall improved, challenges with Class 3 highlight the need for advanced strategies such as specialized augmentations.

C. Conclusion

Unfreezing two blocks enhanced differentiation for minority classes while maintaining calibration reliability. Persistent challenges with Class 3 indicate the need for tailored interventions to further optimize performance.

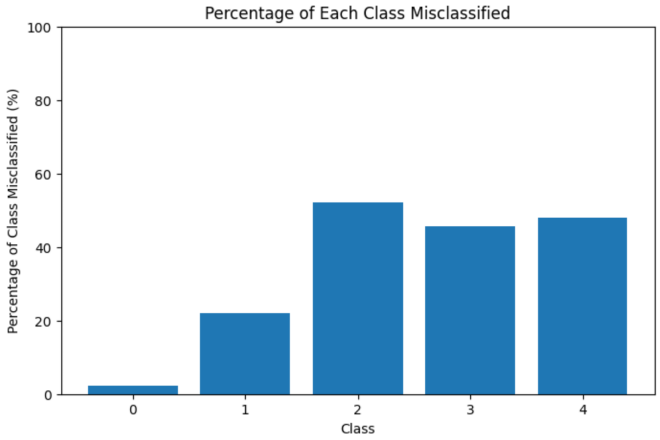

Fig. 5. Misclassification rates per class with class weights and unfreezing two blocks. Improvements are evident for Classes 1 and 4.

VIII. EfficientNet-B4

This experiment evaluated **EfficientNet-B4**, a more complex architecture, to assess its impact on performance. Images were resized to **380x380**, and training strategies from EfficientNet-B0, including class weights, were retained.

A. Key Observations

- **Balanced Accuracy**: Slight improvement over EfficientNet-B0, with marginal class-level gains.
- **Calibration**: Enhanced ECE (**0.0439**) indicated better probability reliability.
- **Class-Specific Challenges**:

  - **Class 2**: Misclassification persisted with steady performance.
  - **Classes 3 and 4**: Minor improvement in Class 4, but Class 3 remained challenging.

B. Trade-offs

- **Increased Training Time**: From **10 minutes (EfficientNet-B0)** to **50 minutes (EfficientNet-B4)**.
- **Limited Performance Gains**: Marginal improvements did not justify the significant increase in complexity.

C. Conclusion

EfficientNet-B4 provided slight enhancements in balanced accuracy and calibration. However, the higher computational cost and continued struggles with minority classes, particularly Class 3, affirm EfficientNet-B0 as the more practical and balanced choice for this dataset.

IX. Advanced Augmentation and FocalLoss

Advanced augmentation, FocalLoss, and their combination were evaluated to address class imbalance and improve generalization. Performance across all three approaches was nearly identical.

A. Key Observations

- **Balanced Accuracy**: Comparable to class weighting with unfreezing two blocks, with no significant improvement.
- **Class-Specific Insights**: Slight gains for Classes 3 and 4, but recall and F1-scores remained suboptimal. Class 0 consistently performed well.
- **Calibration**: ECE slightly worsened (**0.0356**), reducing prediction reliability.

B. Limitations

- Marginal improvements did not justify the added complexity.
- FocalLoss improved minority class focus but reduced calibration.
- CutMix was avoided to preserve subtle features for distinguishing Classes 1 and 2.

X. Results and Discussion

This section summarizes the performance of the different experimental setups.

A. Baseline Model

The baseline achieved a Balanced Accuracy of 56.5%, Kappa of 83.4%, and Accuracy of 78% across five folds, with low Expected Calibration Error (ECE: 0.029). Minority classes (e.g., Class 3 and 4) faced high misclassification rates (50%).

B. Baseline + Class Weight + Unfreeze 2 Blocks

Unfreezing two backbone blocks improved Balanced Accuracy to 64% and Kappa to 84%, though minority class misclassification, especially for Class 3, remained a challenge.

C. EfficientNet-B4

EfficientNet-B4 achieved a Balanced Accuracy of 65% and Accuracy of 76%, with slight performance gains at a higher computational cost, limiting deployment practicality.

D. Advanced Data Augmentation and Focal Loss

This approach yielded a Balanced Accuracy of 63% and Accuracy of 75%. Minority class performance improved modestly, though ECE slightly worsened (0.035).

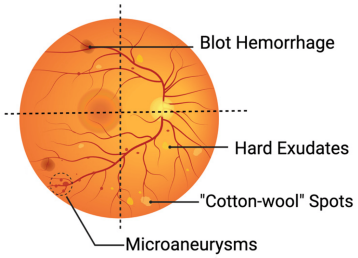

Fig. 6. High inter-class similarity errors between Classes 1 and 2 are often attributed to the subtle presence of ”cotton-wool” spots [2]. The image is taken from [3].

XI. Conclusion

Diabetic retinopathy classification presents challenges stemming from imbalanced datasets and subtle inter-class similarities (e.g., Classes 1 vs. 2 and 3 vs. 4). This study explored various approaches and identified a practical and efficient solution using EfficientNet-B0.

By leveraging essential preprocessing, minimal augmentations, class weighting, and unfreezing two backbone blocks, the model achieved interesting results while preserving subtle features without synthetic data. EfficientNet-B0 balanced computational efficiency and reliability, making it a strong choice for this dataset.

Comparison with public implementations highlighted differences in evaluation strategies. Many focus on overall accuracy, prioritizing dominant classes (e.g., Classes 0 and 2) while neglecting minority classes, inflating metrics like accuracy. This report emphasized balanced metrics and computational efficiency, underscoring the importance of tailored strategies for medical imaging tasks.

References

[1] ”Be careful what you train on,” Kaggle Notebook. Available: https://www.kaggle.com/code/taindow/be-careful-what-you-train-on

[2] ”APTOS : Eye Preprocessing in Diabetic Retinopathy” Kaggle Notebook. Available: https://www.kaggle.com/code/ratthachat/aptos-eye-preprocessing-in-diabetic-retinopathy

[3] Zhao Y, Chen Y, Yan N. The Role of Natural Products in Diabetic Retinopathy. Biomedicines. 2024; 12(6):1138.https://doi.org/10.3390/biomedicines12061138