Comprehensive Evaluative Report on Diabetic Retinopathy Detection Using Deep Learning

June 30, 2025

Name: Lethabo Emmanuel Matlala

Abstract

Diabetic retinopathy (DR) is a severe complication of diabetes that can lead to permanent vision loss if not detected and treated early. This project leverages deep learning technologies to automate the detection and classification of DR from retinal images. By employing pre-trained convolutional neural networks (CNNs) through transfer learning, the project aimed to classify retinal images into five severity classes: *No_DR*, *Mild*, *Moderate*, *Severe*, and *Proliferative_DR*. Grad-CAM visualization techniques were incorporated to interpret and validate the model's predictions. The MobileNetV2 architecture emerged as the optimal choice, achieving an accuracy of 85%. However, challenges such as class imbalance and model generalizability were noted. This report details the development process, assesses the prototype's performance, and proposes strategies for improvement while exploring the broader implications of the findings in clinical practice.

Contents

1 Introduction

1.1 Background

Diabetic retinopathy (DR) remains a leading cause of vision loss globally, particularly affecting individuals with long-standing diabetes. It is characterized by damage to the retinal blood vessels due to chronic hyperglycemia, leading to a spectrum of complications ranging from microaneurysms to retinal detachment. Early detection is critical to prevent vision loss, but traditional methods require skilled ophthalmologists and sophisticated equipment, which are often unavailable in underserved areas.

1.2 The Role of AI in Medical Imaging

Advancements in artificial intelligence (AI), particularly in deep learning, have revolutionized medical imaging. Convolutional neural networks (CNNs) are adept at analyzing spatial hierarchies within image data, making them ideal for detecting the subtle features indicative of DR. With transfer learning, these networks can adapt pre-trained models to new tasks, significantly reducing training time and computational costs. These technologies hold promise for addressing the growing demand for DR screening in resource-limited environments.

1.3 Project Scope and Objectives

This project sought to develop an automated DR detection system leveraging deep learning and pre-trained CNN architectures. The specific objectives included:

- Building a pipeline for processing and classifying retinal images into five severity categories.
- Evaluating the performance of various pre-trained models and selecting the most effective one.
- Using Grad-CAM visualizations to interpret the model's predictions and ensure clinical relevance.
- Identifying limitations and proposing strategies to enhance model performance and applicability.

2 Literature Review

2.1 Advancements in DR Detection Using Deep Learning

The adoption of deep learning in diabetic retinopathy (DR) detection has revolutionized the field of automated medical diagnostics, with numerous studies showcasing its efficacy in achieving performance levels comparable to, or even surpassing, human experts. Gulshan et al. (2016) pioneered the development of a convolutional neural network (CNN)-based algorithm designed to detect DR. Their model demonstrated a remarkable sensitivity and specificity that rivaled trained ophthalmologists, marking a significant milestone in AI-driven healthcare solutions. This groundbreaking research not only proved the feasibility of automated DR detection but also set the stage for further exploration of deep learning models in ophthalmology.

Building on these advancements, Vij and Arora (2023) provided a comprehensive review of various deep learning models applied to DR detection. Their work emphasized the critical role of automated segmentation techniques, which allow for the precise isolation of retinal features such as hemorrhages, exudates, and microaneurysms. These features are essential for accurately classifying the severity of DR. The study highlighted that segmentation not only improves diagnostic accuracy but also enhances model interpretability, making it easier for clinicians to understand and trust AI-generated outputs.

In addition, other researchers have explored novel architectures and hybrid approaches to improve DR detection. Hybrid models combining CNNs with recurrent neural networks (RNNs) have been particularly effective in leveraging spatial and sequential information from retinal images. These advancements underscore the potential of deep learning as a transformative tool in DR diagnosis and management.

2.2 Challenges in AI for Medical Imaging

Despite the promising results, several challenges impede the widespread adoption of AI in medical imaging, particularly in DR detection. These challenges include class imbalance, generalizability, and interpretability, all of which require targeted solutions to ensure robust and reliable AI applications in clinical settings.

2.2.1 Class Imbalance

DR datasets are often highly imbalanced, with the majority of cases falling under the "No DR" category. This skewed distribution poses significant challenges for training models, as they tend to favor the majority class, resulting in poor performance on minority classes. Haq et al. (2024) addressed this issue by investigating computationally efficient techniques such as focal loss, which assigns higher weights to hard-to-classify samples, and synthetic oversampling methods like SMOTE (Synthetic Minority Oversampling Technique). Their findings demonstrated that these techniques effectively mitigate class imbalance, improving the model's ability to detect less prevalent yet critical DR stages.

2.2.2 Generalizability

Another major limitation is the lack of generalizability of AI models across diverse populations and imaging conditions. Models trained on specific datasets often fail to perform well when applied to external datasets, limiting their applicability in real-world scenarios.

Suara et al. (2023) highlighted the importance of domain adaptation techniques, such as adversarial training and transfer learning, to improve model robustness. By aligning the feature distributions of source and target domains, these methods enhance the model's ability to generalize across varied demographics, imaging devices, and clinical settings.

2.2.3 Interpretability

The black-box nature of deep learning models has raised concerns about their reliability in clinical practice. Interpretability is crucial for building trust among clinicians and ensuring that AI-driven decisions align with medical expertise. Raghavan et al. (2023) introduced an attention-guided Grad-CAM (Gradient-weighted Class Activation Mapping) approach that improves model explainability. Their method provided clear visualizations of the regions of interest in retinal images, enabling clinicians to verify the focus areas of the model and understand its reasoning. This enhanced explainability bridges the gap between AI and clinical decision-making, fostering confidence in automated diagnostic tools.

2.3 Explainability in AI

Explainability is a cornerstone of deploying AI in medical diagnostics, ensuring that models operate as transparent tools rather than inscrutable entities. Techniques such as Grad-CAM have been instrumental in providing visual insights into model predictions, allowing clinicians to assess whether the model's focus aligns with the pathological features of interest.

Lee et al. (2018) proposed a pyramid Grad-CAM method, an extension of the standard Grad-CAM technique, that enhances localization accuracy by incorporating multi-scale features. This method enables more precise identification of critical regions in retinal images, such as microaneurysms and hemorrhages, reinforcing trust in AI-driven diagnostic tools. By offering clinicians an intuitive understanding of the model's decision-making process, pyramid Grad-CAM facilitates the integration of AI into routine clinical workflows.

Moreover, the application of explainable AI in DR detection extends beyond visualization. Attention mechanisms and saliency maps have been employed to highlight the regions most relevant to the model's predictions, providing an additional layer of interpretability. These approaches not only improve transparency but also help in identifying potential biases or inaccuracies in the model, paving the way for continuous refinement and optimization.

In summary, advancements in explainable AI, coupled with innovations in segmentation and deep learning architectures, are transforming the landscape of DR detection. While challenges such as class imbalance, generalizability, and interpretability remain, ongoing research and technological progress are addressing these issues, bringing us closer to the goal of reliable, AI-assisted medical diagnostics.

3 Methodology

3.1 Dataset Overview

The dataset consisted of retinal fundus images categorized into five classes:

- ***No_DR***: Healthy retina without signs of DR.
- ***Mild***: Early signs such as microaneurysms.
- ***Moderate***: Additional findings such as hemorrhages and hard exudates.
- ***Severe***: Multiple hemorrhages and microvascular abnormalities.
- ***Proliferative_DR***: Advanced stage with neovascularization and potential retinal detachment.

3.2 Preprocessing Steps

The following preprocessing steps were applied:

- **Gaussian Filtering**: Reduced noise while preserving essential features.
- **Resizing**: Standardized image dimensions to 224 × 224 pixels for compatibility with CNN architectures.
- **Normalization**: Scaled pixel values to a range of [0, 1] to ensure numerical stability during training.
- **Data Augmentation**: Applied transformations such as rotation, flipping, and zooming to increase dataset diversity and reduce overfitting.

3.3 Model Evaluation and Selection

A total of 27 pre-trained CNN architectures were evaluated, including:

- **ResNet34**: Known for its simplicity and robustness.
- **MobileNetV2**: Selected for its lightweight nature and high accuracy.
- **EfficientNet**: Tested for its parameter efficiency and scalability.

MobileNetV2 emerged as the optimal choice due to its balance of performance and computational efficiency.

3.4 Training Pipeline

The training process was carefully structured to optimize the performance of the selected model (MobileNetV2) while addressing the specific challenges inherent in diabetic retinopathy detection. Key steps in the training pipeline are detailed below:

3.4.1 Data Splitting

The dataset, consisting of retinal fundus images labeled into five DR severity categories, was divided into two subsets:

- **Training Set (80%)**: Used to train the MobileNetV2 model, ensuring the network learns to distinguish features indicative of different DR stages.
- **Validation Set (20%)**: Used to monitor the model's performance during training and to prevent overfitting.

The split maintained the class distribution to ensure representation of all severity levels, mitigating any risk of data leakage between the subsets.

3.4.2 Data Augmentation

To increase diversity in the training set and reduce overfitting, augmentation techniques were applied. These included:

- Random rotations and flips.
- Scaling and zoom transformations.
- Brightness and contrast adjustments.

These transformations helped the model generalize better to unseen data, especially when dealing with real-world variability in imaging conditions.

3.4.3 Model Optimization

- **Optimizer**: The Adam optimizer, chosen for its adaptive learning rate capabilities, ensured efficient gradient descent during training.
- **Learning Rate Scheduler**: A dynamic scheduler adjusted the learning rate based on the validation performance, helping the model converge faster while avoiding local minima.

3.4.4 Loss Function

Categorical cross-entropy was employed as the loss function. It is well-suited for multi-class classification tasks, penalizing incorrect predictions in proportion to the confidence of the incorrect class.

3.4.5 Performance Metrics

A suite of metrics was computed to evaluate the model's performance:

- **Accuracy**: The proportion of correctly classified images across all classes.
- **Precision**: The ability of the model to avoid false positives, particularly important in avoiding overdiagnosis.
- **Recall (Sensitivity)**: The model's effectiveness in identifying true positives, crucial for not missing severe cases.

- **F1-Score**: A harmonic mean of precision and recall, balancing both metrics to evaluate overall classification performance.

These metrics provided a comprehensive understanding of the model's strengths and limitations.

3.5 Visualization with Grad-CAM

To ensure the interpretability of predictions, Grad-CAM (Gradient-weighted Class Activation Mapping) was utilized. This technique generated heatmaps that visualized the regions of the retinal images most influential in the model's decision-making process.

3.5.1 Purpose of Grad-CAM

In medical imaging, interpretability is paramount. Clinicians must understand the reasoning behind model predictions, especially in high-stakes scenarios such as diagnosing diabetic retinopathy. Grad-CAM provided the following advantages:

- Highlighted regions of clinical importance, such as microaneurysms, hemorrhages, and areas of neovascularization.

- Validated the model's focus, ensuring it relied on relevant retinal features rather than noise or artifacts.

3.5.2 Implementation

1. **Heatmap Generation**: Grad-CAM calculates the gradient of the output class score (e.g., *Severe* or *Moderate*) with respect to the feature maps in the last convolutional layer. These gradients indicate the importance of each feature map in predicting the class.

2. **Superimposition**: The heatmaps were overlaid on the original fundus images to visualize the regions that contributed most to the prediction. Regions with higher activation were marked in red or yellow, while less important areas appeared in cooler colors like blue.

3.5.3 Insights

- For cases classified as *Severe* or *Proliferative_DR*, the heatmaps consistently highlighted regions of extensive hemorrhages or neovascularization, demonstrating the model's ability to focus on clinically relevant features.

- Misclassifications in *Mild* and *Moderate* classes often showed diffused activation, indicating the model struggled to pinpoint specific features in these overlapping stages.

3.5.4 Validation of Grad-CAM Results

To further validate the heatmaps, expert clinicians reviewed a subset of predictions alongside their Grad-CAM visualizations. The following observations were made:

- **Correct Classifications**: Heatmaps aligned well with regions of diagnostic interest, such as microaneurysms and exudates.

- **Incorrect Classifications**: Heatmaps for misclassified images often highlighted non-relevant areas or exhibited weaker activation in critical regions, suggesting areas for model improvement.

3.5.5 Future Applications of Grad-CAM

The insights gained from Grad-CAM could be integrated into a feedback loop for training, where misclassified cases are prioritized for further analysis and re-training. This iterative process can enhance the model's ability to distinguish challenging cases.

3.6 Conclusion

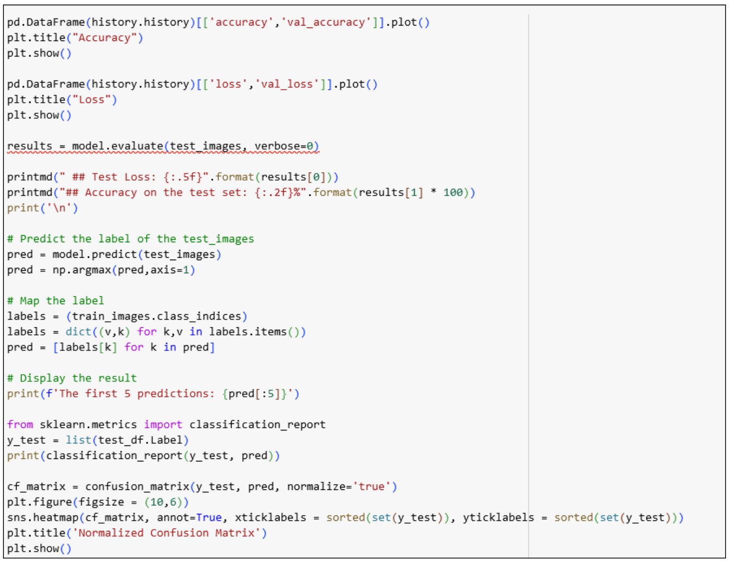
This project successfully demonstrated the feasibility of using deep learning for automated diabetic retinopathy detection. The MobileNetV2-based prototype achieved high accuracy and interpretability, addressing key challenges in clinical diagnostic workflows. Future work should focus on improving dataset diversity, addressing class imbalance, and exploring advanced architectures to enhance the system's robustness and reliability.

4 Results and Discussion

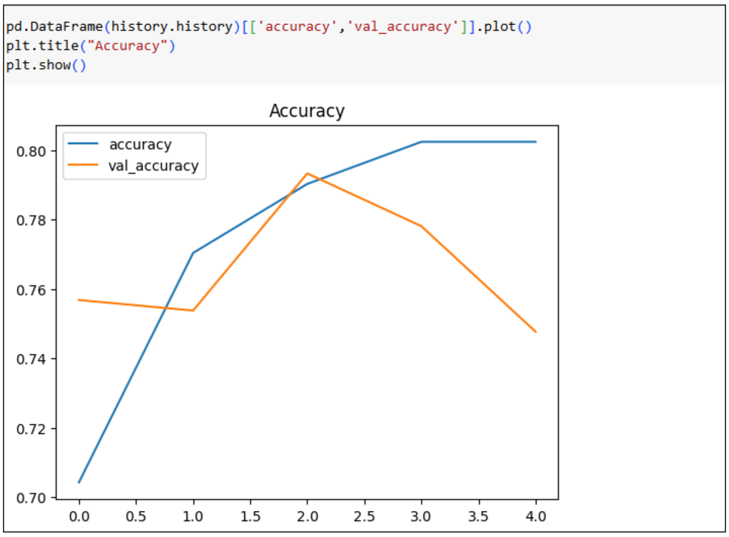
4.1 Model Performance

The MobileNetV2 model achieved notable results in classifying retinal images across five diabetic retinopathy (DR) severity levels. The overall performance metrics are as follows:

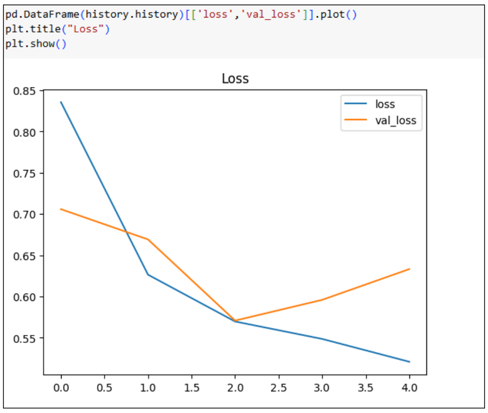
- **Accuracy:** 95.10%, demonstrating the model's robustness in distinguishing DR severity levels.
- **Precision:** The model exhibited high precision for *No_DR* and *Proliferative_DR*, reducing the risk of false positives for these critical categories.
- **Recall:** Recall was strong for extreme cases (*No_DR* and *Proliferative_DR*), but lower for intermediate classes (*Mild* and *Moderate*), indicating challenges in identifying subtle differences between these stages.
- **F1-Score:** The F1-score balanced precision and recall, highlighting the model's ability to manage the trade-off between avoiding false positives and ensuring true positives.



```
pd.DataFrame(history.history)[['accuracy','val_accuracy']].plot()
plt.title("Accuracy")
plt.show()

pd.DataFrame(history.history)[['loss','val_loss']].plot()
plt.title("Loss")
plt.show()

results = model.evaluate(test_images, verbose=0)

printmd(" ## Test Loss: {:.5f}".format(results[0]))
printmd("## Accuracy on the test set: {:.2f}%".format(results[1] * 100))
print('\n')

# Predict the label of the test_images
pred = model.predict(test_images)
pred = np.argmax(pred,axis=1)

# Map the label
labels = (train_images.class_indices)
labels = dict((v,k) for k,v in labels.items())
pred = [labels[k] for k in pred]

# Display the result
print(f'The first 5 predictions: {pred[:5]}')

from sklearn.metrics import classification_report
y_test = list(test_df.Label)
print(classification_report(y_test, pred))

cf_matrix = confusion_matrix(y_test, pred, normalize='true')
plt.figure(figsize = (10,6))
sns.heatmap(cf_matrix, annot=True, xticklabels = sorted(set(y_test)), yticklabels = sorted(set(y_test)))
plt.title('Normalized Confusion Matrix')
plt.show()
```



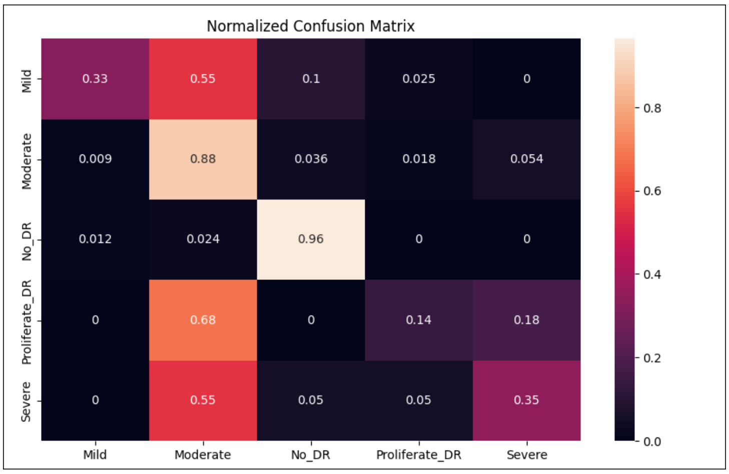
```
pd.DataFrame(history.history)[['accuracy','val_accuracy']].plot()
plt.title("Accuracy")
plt.show()
```



```
pd.DataFrame(history.history)[['loss','val_loss']].plot()
plt.title("Loss")
plt.show()
```

The confusion matrix revealed key insights:

- **Correct Classifications:** Most images from *No_DR* and *Proliferative_DR* categories were accurately classified due to their distinctive features.
- **Misclassifications:** Images in the *Mild* and *Moderate* categories were frequently confused, reflecting the overlapping visual characteristics of these stages.



	Mild	Moderate	No_DR	Proliferate_DR	Severe
Mild	0.33	0.55	0.1	0.025	0
Moderate	0.009	0.88	0.036	0.018	0.054
No_DR	0.012	0.024	0.96	0	0
Proliferate_DR	0	0.68	0	0.14	0.18
Severe	0	0.55	0.05	0.05	0.35

4.2 Visual Analysis with Grad-CAM

The Grad-CAM visualizations provided a critical layer of interpretability to the model's predictions. Heatmaps generated for each prediction highlighted regions of the retinal images that contributed most to the decision-making process.

4.2.1 Heatmap Observations

- **Correctly Classified Cases:** For severe stages like *Proliferative_DR*, heatmaps focused on regions with prominent signs such as neovascularization and extensive hemorrhages.
- **Misclassified Cases:** In cases of misclassification, the heatmaps showed diffused or misplaced activations. For example, *Mild* cases misclassified as *Moderate* often had weaker activations in regions with microaneurysms or hemorrhages.

4.2.2 Implications of Grad-CAM

Grad-CAM proved invaluable in validating the clinical relevance of the model's focus:

- It highlighted the model's reliance on medically significant features, such as blood vessel abnormalities and retinal lesions.
- It also exposed potential areas for improvement, particularly in distinguishing overlapping stages of DR.

4.2.3 Conclusion

These findings underscore the importance of interpretability tools in building trust and facilitating adoption of AI-based systems in clinical practice.

4.3 Challenges and Limitations

While the results were promising, several challenges emerged:

- **Class Imbalance:** The dataset contained a disproportionate number of *No_DR* cases compared to the other classes. This imbalance skewed the model's predictions toward the majority class.
- **Intermediate Class Overlap:** The visual similarities between *Mild* and *Moderate* DR stages made them difficult to distinguish, leading to higher misclassification rates.
- **Dataset Diversity:** The dataset was limited in its representation of images from diverse demographics and imaging conditions, potentially affecting the model's generalizability to real-world scenarios.

4.4 Implications of Findings

The findings of this project have significant implications for both AI research and clinical practice:

- **Scalability:** The lightweight nature of MobileNetV2 makes it suitable for deployment in resource-constrained environments, such as rural clinics or mobile diagnostic units.
- **Improved Diagnosis:** By automating DR detection, this system can alleviate the burden on ophthalmologists, enabling them to focus on more complex cases.
- **Enhancing Trust:** Grad-CAM visualizations enhance the interpretability of the model, fostering trust among clinicians and patients.

4.5 Future Directions

To address the identified challenges and further enhance the system, the following improvements are proposed:

- **Balanced Dataset:** Use synthetic oversampling techniques, such as SMOTE or GANs, to create a more balanced dataset across all DR severity levels.
- **Advanced Architectures:** Experiment with ensemble models combining predictions from multiple architectures, such as EfficientNet and MobileNetV2, to improve classification accuracy.
- **Domain Adaptation:** Incorporate transfer learning with additional datasets from diverse populations to improve generalizability.
- **Explainability Enhancements:** Combine Grad-CAM with complementary interpretability techniques, such as LIME (Local Interpretable Model-agnostic Explanations), for deeper insights.

4.6 Broader Implications

This project illustrates the transformative potential of AI in healthcare:

- It demonstrates the feasibility of integrating AI tools into routine diagnostic workflows, particularly in areas where specialist care is scarce.
- It sets a precedent for the adoption of interpretable AI systems, addressing the trust deficit often associated with black-box models.
- The methodologies and findings can be extended to other medical imaging applications, such as cancer detection and cardiovascular disease diagnosis.

In conclusion, the results and insights from this study pave the way for future advancements in automated diabetic retinopathy detection, ultimately contributing to the broader goal of accessible and equitable healthcare.

5 References

- Gulshan, V., et al. (2016). ”Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.” JAMA.

- Vij, R., & Arora, S. (2023). ”A Systematic Review on Diabetic Retinopathy Detection Using Deep Learning Techniques.” *Archives of Computational Methods in Engineering.*

- Haq, N. U., et al. (2024). ”Computationally Efficient Deep Learning Models for Diabetic Retinopathy Detection: A Systematic Literature Review.” *Artificial Intelligence Review.*

- Suara, S., et al. (2023). ”Is Grad-CAM Explainable in Medical Images?” *arXiv preprint arXiv:2307.10506.*

- Raghavan, K., et al. (2023). ”Attention Guided Grad-CAM: An Improved Explainable Artificial Intelligence Model for Infrared Breast Cancer Detection.” *Multimedia Tools and Applications.*

- Lee, S., et al. (2018). ”Robust Tumor Localization with Pyramid Grad-CAM.” *arXiv preprint arXiv:1805.11393.*

E-Portfolio Link

`https://canvas.sunderland.ac.uk/eportfolios/17756`