IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Explainable And Temporal Attention Analysis For Deep Learning-Based Diabetic Retinopathy Severity Grading

Interpretability and Ablation Study of a Multi-Scale CNN-RNN Architecture

[1]Rahul Yadav
[1]M.Tech. Student
[1]Department of Computer Science and Engineering,
[1]Goel Institute of Technology & Management, Lucknow, India

Abstract: Deep learning systems for diabetic retinopathy (DR) severity grading have achieved diagnostic accuracy comparable to expert clinicians, yet their adoption is limited by the opacity of their internal decision-making. This paper presents an interpretability- and ablation-focused analysis of a previously developed hybrid DR classification architecture that combines a DenseNet-121 backbone with a Feature Pyramid Network (FPN), a Bidirectional LSTM (Bi-LSTM) for temporal modelling of sequential retinal examinations, and dual-level spatial and temporal attention mechanisms. Rather than re-reporting overall classification accuracy, this work examines the model interpretability components - Grad-CAM activation maps, scale-specific spatial attention, and temporal attention weight visualisation - and quantifies the individual contribution of multi-scale feature extraction, temporal attention, and the composite loss function through systematic ablation experiments. On the Kaggle Diabetic Retinopathy Detection dataset, removing the FPN reduced the macro-AUC from 0.96 to 0.94, removing temporal attention reduced overall accuracy by approximately 1.2%, and removing the focal-loss component increased false negatives in minority severity classes, with the composite loss yielding a macro-F1 improvement of approximately 2%. Grad-CAM visualisations confirmed that model attention consistently localises microaneurysms, haemorrhages, and exudates relevant to each severity grade. These results demonstrate that the architectural and training choices made in the base model are individually justified and that the resulting attention maps provide an interpretable basis for clinical trust in AI-assisted DR screening.

Index Terms - Diabetic Retinopathy, Explainable AI, Grad-CAM, Attention Mechanism, Feature Pyramid Network, Bi-LSTM, Ablation Study, Deep Learning

I. Introduction

Diabetic retinopathy (DR) is a progressive complication of diabetes mellitus and remains a leading cause of preventable blindness in the working-age population worldwide. The International Diabetes Federation estimated approximately 537 million adults living with diabetes in 2021, with projections of around 783 million by 2045, and the World Health Organization attributes approximately 4.8% of global blindness cases to DR. Conventional DR screening relies on manual evaluation of colour fundus photographs by trained ophthalmologists, a process that is resource-intensive, subject to inter-rater variability, and unable to scale with the growing diabetic population.

In a previous study, the authors developed a hybrid deep learning architecture for five-class DR severity grading that integrates a DenseNet-121 backbone, a Feature Pyramid Network (FPN) for multi-scale lesion representation, and a Bidirectional LSTM (Bi-LSTM) with temporal attention for modelling longitudinal disease progression across sequential clinic visits. That architecture achieved an overall accuracy of 92.4%, a macro-averaged F1-score of 0.90, a Cohen's Kappa of 0.89, and a macro-AUC of 0.96 on the Kaggle Diabetic Retinopathy Detection dataset, outperforming single-branch CNN baselines such as ResNet50 and DenseNet121.

While these results establish the predictive viability of the architecture, two questions remain insufficiently addressed: (i) what is the individual contribution of each architectural and training component to the reported performance, and (ii) does the model's internal attention correspond to clinically meaningful retinal biomarkers. This paper addresses both questions through a structured ablation study and an interpretability analysis based on Grad-CAM and attention-weight visualisation, both of which were performed as part of the original experimental work but not previously analysed in depth.

II. Related Work

Early automated DR screening relied on handcrafted feature extraction - morphological top-hat filtering and scale-space Gaussian methods for microaneurysm detection, matched filters and Gabor filter banks for vessel segmentation, and region-growing algorithms for exudate and haemorrhage localisation - followed by classical classifiers such as support vector machines, k-nearest neighbours, and ensemble trees. These methods were interpretable by construction but suffered from limited scalability and poor generalisation across imaging conditions, and struggled to distinguish visually similar adjacent severity grades such as mild and moderate non-proliferative DR (NPDR).

The landmark study of Gulshan et al. demonstrated that deep convolutional neural networks trained on over 100,000 expert-annotated fundus photographs could exceed 90% sensitivity and specificity for referable DR. Subsequent work explored ResNet, DenseNet, EfficientNet, and Inception-based backbones, transfer learning from ImageNet, and multi-task learning frameworks combining DR grading with image-quality assessment. Hybrid CNN-RNN architectures incorporating Bi-LSTM components were later proposed to exploit the longitudinal trajectory of DR progression, with reported accuracy gains of up to 94% on sequential datasets.

In parallel, attention mechanisms - including Squeeze-and-Excitation channel recalibration, the Convolutional Block Attention Module, and self-attention in vision transformers - and gradient-based visualisation techniques such as Grad-CAM and Score-CAM have been used to improve both accuracy and interpretability. However, the literature commonly reports these components as standalone additions without a systematic ablation that isolates the contribution of each, motivating the analysis presented in this paper.

III. Methodology

A. Base Architecture

The architecture analysed in this paper follows a five-stage pipeline: (1) data acquisition and preprocessing, (2) CNN-based feature extraction using a DenseNet-121 backbone adapted for 512x512 RGB fundus images, (3) multi-scale feature extraction via a Feature Pyramid Network producing four feature levels P1-P4 with channel and spatial attention sub-modules, (4) temporal modelling of sequential fundus images using a two-layer Bi-LSTM (hidden size 256 per direction) with a feed-forward temporal attention sub-network that produces softmax-weighted context vectors, and (5) a classification head with a 256-unit fully connected layer, dropout of 0.5, and a five-class softmax output corresponding to No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and PDR.

Preprocessing included blur detection via Laplacian variance, CIELAB colour normalisation with L-channel histogram equalisation, CLAHE-based contrast enhancement, and data augmentation comprising geometric transformations, colour jittering, noise injection, and Mixup/CutMix. Training used the Adam optimiser with an initial learning rate of 1e-4, cosine annealing with a five-epoch warm-up, L2 regularisation of 1e-5, gradient clipping, a batch size of 32, mixed-precision training, early stopping with a patience of 15 epochs, and five-fold cross-validation.

B. Composite Loss Formulation

To counteract the pronounced class imbalance present in the Kaggle and EyePACS datasets, a composite loss function was used, combining weighted cross-entropy loss, focal loss, and auxiliary losses computed at intermediate FPN outputs:

$$L = \alpha \cdot L\_CE + \beta \cdot L\_FL + \gamma \cdot L\_Aux$$

where the scalar coefficients α, β, and γ govern the relative contribution of each term and were determined through validation-set hyperparameter search.

C. Interpretability Tools

Three complementary interpretability mechanisms were evaluated: (i) Grad-CAM and guided backpropagation, applied at the final convolutional layer of the DenseNet-121 backbone to generate class-discriminative heatmaps; (ii) scale-specific spatial and channel attention maps from each FPN level (P1-P4), visualising which spatial scale the model emphasises for a given severity grade; and (iii) temporal attention weight plots over the Bi-LSTM hidden states, indicating which clinical visit(s) most influenced the final prediction in a longitudinal sequence.

D. Ablation Design

Three ablation experiments were conducted on the Kaggle Diabetic Retinopathy Detection dataset, each removing a single architectural or training component while keeping all other settings identical to the base configuration: (1) removal of the Feature Pyramid Network, retaining only the DenseNet-121 backbone output; (2) removal of the temporal attention sub-network, using the final Bi-LSTM hidden state directly for classification; and (3) removal of the focal-loss term from the composite loss, retaining only weighted cross-entropy and auxiliary FPN losses.

IV. Results and Discussion

A. Multi-Scale Feature Extraction (FPN Ablation)

Removing the FPN and relying solely on the DenseNet-121 backbone reduced the macro-AUC from 0.96 to 0.94, corresponding to an overall accuracy reduction of approximately 1.5%. This confirms that lesions of different physical scales - microaneurysms at one extreme and extensive haemorrhages or neovascularisation at the other - benefit from the hierarchical multi-resolution representation provided by the FPN, consistent with the broad spectrum of lesion sizes characteristic of DR pathology.

Configuration	Macro-AUC	Approx. Accuracy Change
With FPN (proposed)	0.96	Baseline
Without FPN	0.94	-1.5%

Table I: Effect of Multi-Scale Feature Extraction (FPN) on Performance

B. Temporal Attention (Bi-LSTM Ablation)

Table II summarises the effect of the temporal attention sub-network on overall classification accuracy. Without temporal attention, overall accuracy decreased by approximately 1.2%, and qualitative inspection of the temporal attention weight plots (Figure 4.5 of the underlying study) showed that the attention mechanism reliably identifies the clinical visit(s) most informative for severity grading, providing an interpretable trace of which examination time-points drove a given prediction. This is particularly relevant for early-stage NPDR, where disease progression across visits carries diagnostic signal not present in any single image.

Configuration	**Effect on Overall Accuracy**	**Interpretability Outcome**
With temporal attention (proposed)	Baseline (92.4%)	Visit-level attention weights identify the most informative examination(s)
Without temporal attention	Approx. -1.2%	No visit-level interpretability; final hidden state used directly

Table II: Effect of Temporal Attention on Accuracy and Interpretability

C. Composite Loss Function

Removing the focal-loss component from the composite loss increased false negatives in minority severity classes (Mild and Severe NPDR), the categories most underrepresented in the Kaggle and EyePACS datasets. Restoring the composite loss formulation, combining weighted cross-entropy, focal loss, and auxiliary FPN losses, improved the macro-averaged F1-score by approximately 2%, demonstrating that explicit handling of class imbalance is necessary to maintain sensitivity for the clinically critical early-stage categories.

D. Grad-CAM and Attention-Based Interpretability

Grad-CAM heatmaps generated across all five DR severity classes showed a consistent qualitative pattern: for mild DR cases, activation was concentrated on small, localised regions corresponding to microaneurysms; for severe and proliferative DR cases, activation extended over larger regions corresponding to extensive haemorrhages and areas of neovascularisation. Misclassification analysis indicated that mild DR was most frequently confused with the no-DR class, owing to the subtlety of early lesions, while severe-versus-PDR misclassifications occurred predominantly on lower-quality images affected by illumination artefacts.

Combined with the scale-specific FPN attention and the temporal attention weight plots, these visualisations provide a multi-level interpretability stack: spatial attention indicates which retinal scale the model is reasoning at, Grad-CAM indicates which retinal regions drive the prediction, and temporal attention indicates which examination visit(s) were most influential. This layered interpretability directly addresses the model-transparency concern identified as a key barrier to clinical adoption of AI-assisted DR screening.

E. Computational Considerations

The full architecture (DenseNet-121 + FPN + Bi-LSTM + attention) comprises approximately 14 million parameters, substantially fewer than larger backbones such as ResNet-152 (60 million parameters), while achieving a mean inference time of approximately 90-100 ms per 512x512 image on an NVIDIA RTX 3090 GPU and approximately 750 ms on an Intel Core i9-11900K CPU. A quantised version of the model retains approximately 6 million parameters with minimal accuracy loss, supporting deployment in resource-constrained screening environments without compromising the interpretability components described above.

V. Conclusion

This paper presented an interpretability- and ablation-focused analysis of a hybrid DenseNet-121, FPN, and Bi-LSTM architecture with dual-level attention for diabetic retinopathy severity grading. Systematic ablation confirmed that multi-scale feature extraction (FPN), temporal attention, and the composite loss function each make a measurable, individually justified contribution to model performance, with the FPN improving macro-AUC by 0.02, temporal attention improving overall accuracy by approximately 1.2%, and the composite loss improving macro-F1 by approximately 2% relative to ablated configurations. Grad-CAM and attention-weight visualisations further demonstrated that the model's predictions are anchored to clinically relevant biomarkers - microaneurysms, haemorrhages, and exudates - and to diagnostically informative examination visits in longitudinal sequences. Together, these findings strengthen the case for the architectural and training design adopted in the base model and provide a concrete interpretability basis for its use in clinical DR screening workflows. Future work will extend this analysis with quantitative lesion-overlap metrics for Grad-CAM activations and evaluate clinician feedback on the generated attention visualisations.

Acknowledgment

The author expresses sincere gratitude to Dr. Anita Pal, Professor, Department of Computer Science and Engineering, Goel Institute of Technology & Management, Lucknow, for her valuable guidance and support throughout this research. The author also acknowledges Goel Institute of Technology & Management and Dr. APJ Abdul Kalam Technical University, Lucknow, for providing the necessary resources and academic environment.

References

[1] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402-2410, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770-778.

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 4700-4708.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, 2018, pp. 7132-7141.

[5] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618-626.

[6] Z. Xu et al., "Attention-based CNN for diabetic retinopathy severity detection," Biomed. Signal Process. Control, vol. 65, p. 102323, 2021.

[7] L. Zhang et al., "Temporal attention network for longitudinal diabetic retinopathy screening," IEEE Trans. Med. Imaging, vol. 40, no. 9, pp. 2233-2242, 2021.

[8] J. Jabbar, Y. Gao, and S. Hao, "Temporal modeling of diabetic retinopathy using hybrid CNN-RNN architecture," IEEE Access, vol. 12, pp. 14578-14589, 2024.

[9] S. Zhang et al., "Feature pyramid networks with attention for diabetic retinopathy," Med. Image Anal., vol. 66, p. 101813, 2021.

[10] M. Alavee et al., "Explainable AI for diabetic retinopathy," Expert Syst. Appl., vol. 216, p. 119299, 2024.

[11] R. Yadav and A. Pal, "Diabetic Retinopathy Classification Using Multi-Scale and Temporal Attention," Int. J. Creative Research Thoughts (IJCRT), 2026.