

An Explainable Deep Learning Framework for Multi-Disease Ocular Classification and Myopia Severity Estimation

Sharat Acharja, Azahar Hossen Novel, Amikur Rahman

Department of Computer Science and Engineering

Green University of Bangladesh

Dhaka, Bangladesh

sharatacharjee6@gmail.com, rampardnovel83@gmail.com, thamik67@gmail.com

Abstract—Visual impairment caused by eye diseases and refractive errors has been a prominent global health concern. Though Deep Learning has depicted high performance in retinal fundus image analysis, existing methods in this area also lack interpretability and do not deal with disease severity. In this paper, an explainable Deep Learning model has been proposed for multi-disease ocular classification and myopia severity analysis based on retinal fundus images. Transfer Learning has been utilized for discriminative feature extraction, and visualization methods using AI interpretability have been incorporated for visualizing regions important for decision-making. A guideline-based clinical decision-making module has been incorporated for the initial management of myopia. Performance analysis on the ODIR-5K dataset has shown consistency in classification accuracy and interpretability of results, and thus it can act as a supporting tool in detecting ocular diseases and their severity assessments.

The source code is publicly available at: [GitHub Repository](#).

Index Terms—Fundus Image Analysis, Ocular Disease Classification, Myopia Severity Estimation, Explainable Artificial Intelligence, Deep Learning, Clinical Decision Support

I. INTRODUCTION

Vision impairment caused by some ocular diseases and errors of refraction is still a challenge to many. If left untreated, conditions like diabetic retinopathy, glaucoma, cataract, age-related macular degeneration, hypertensive retinopathy, and myopia can result in progressive vision loss. If their conditions are not treated, they may result in progressive vision loss. A popular non-invasive technique for seeing retinal pathology, retinal fundus photography is perfect for automated large-scale screening. Recently, deep learning techniques such as convolutional neural networks have boosted the analysis of the fundus image significantly by extracting discriminative features from the data itself [1], [2]. But these approaches are mostly concerned with the detection or optimization of the screening process of a single disease and disregard the interpretability of the results or the severity of the respective diseases.

Computation related to myopia, a worldwide pervasive refractive problem with an escalating prevalence [3], has largely utilized binary classification. In addition, the estimation

of myopia severity from fundus images, without actually refracting, has not gained adequate attention.

A lack of explainability makes it even more difficult for adoption in the clinical environment. There exist explainable methods of AIs, like gradient visualization, where regions of an image pushing the prediction can be emphasized [4].

This paper introduces an explainable deep learning framework that combines multi-disease ocular recognition and myopia severity assessment on the basis of fundus images. This framework incorporates the process of visual explanations and a decision support system based on a guideline for the initial assessment of myopia. The major contributions of the paper are:

- A explainable framework for multi-disease ophthalmic classification from retinal fundus images. The output is a set of annotations in the form of diagnostic keywords from the input sentence.
- Use of Explainable AI Methods to Facilitate Interpretation.
- A rule-based clinical decision support component for initial myopia treatment.

II. RELATED WORK

Initial analysis of the retinal scans used handmade features and traditional classifiers. Yet these were prone to a lack of generality. Later, convolutional networks facilitated the end-to-end learning process and led to improved approaches to the diagnosis of diabetic retinopathy, glaucoma, cataract, and AMD [5], [6].

Although single-disease diagnosis has made progress, multi-disease classification is relatively under-explored, and most existing work focuses on accuracy rather than modeling severity or interpretability [7]. Research in myopia detection is almost entirely binary classification, with little work on severity estimation, which is generally done through reference to an independent measure rather than image inference.

Interpretability can be considered another large gap that needs to be filled. The reason for that is that most existing models are black-box models, and as a consequence, they are

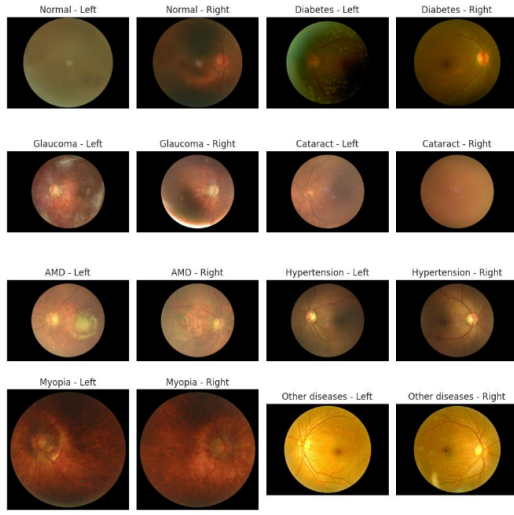


Fig. 1: Representative color fundus images from each ocular disease category in the ODIR-5K dataset.

hardly trusted in any medical setting. Yet, there are explainable methods, for instance, Grad-CAM, utilized in medical images [4], [8]

Unlike existing research, this proposed method puts together multi-disease prediction, estimation of myopia severity, visualization of explanations, and decision support in accordance with guidelines as a whole.

III. METHODOLOGY

A. Dataset Description

In the current research, the Ocular Disease Intelligent Recognition (ODIR-5K) [9] benchmark is considered for identifying the bilateral color fundus images with the diagnostic keywords. This benchmark has been established for eight eye-related diseases and supports the classification of multiple diseases along with the detection of severe myopia with the help of keywords. Only valid retinal images were used, excluding any corrupted files and invalid records.

B. Label Creation and Data Preprocessing

All images have been resized to 224×224 pixels and the intensity value has been normalized. Multi-class labels are created from the diagnosis terms. Myopia is classified into three categories: pathological (-8.0 D to -20.0 D), strong (-3.0 D to -8.0 D), and slight (-0.5 D to -3.0 D). Train, validation, and test data make up the data set.

C. Deep Learning Model Architecture

Transfer learning from a pre-trained ResNet50 model was used. on ImageNet [2]. For this purpose, the convolutional base was frozen and custom classification layers were added: global average pooling is followed by fully connected layers and softmax output. The model was trained with categorical cross-entropy loss and the Adam optimizer.

Explainable Machine Learning Framework for Precision Eye Care

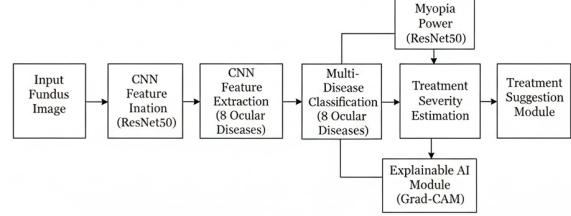


Fig. 2: Overall architecture of the proposed explainable framework showing fundus image input, multi-disease classification, myopia severity estimation, explainable AI, and clinical decision support.

D. Implementation Details and Model Selection

Various architectures were explored in the experiment: ResNet50, VGG16, VGG19, Xception) were under the same conditions. ResNet50 provided the best balance of performance, stability, and explainability integration, and was selected for detailed analysis. and reporting. All experiments were implemented using TensorFlow/Keras; code is publicly available

E. Explainable Artificial Intelligence Integration

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to produce heat maps illustrating regions of the image which had a significant influence on predictions, thus facilitating the validation of anatomic focus.

F. Clinical Decision Support Strategy

The rule-based module is used to translate the severity of the predicted cases of myopia into a set of non-invasive recommendations generated from the guidelines. The recommendations aim at promoting early screening for the condition.

IV. EXPERIMENTAL RESULTS

A. Multi-class ROC Analysis

Variety in performance is observed in class-specific ROC curves. Glaucoma is separable, while Hypertension and Cataract have relatively lower AUC values because of their delicate details and visual similarities and class imbalance, respectively.

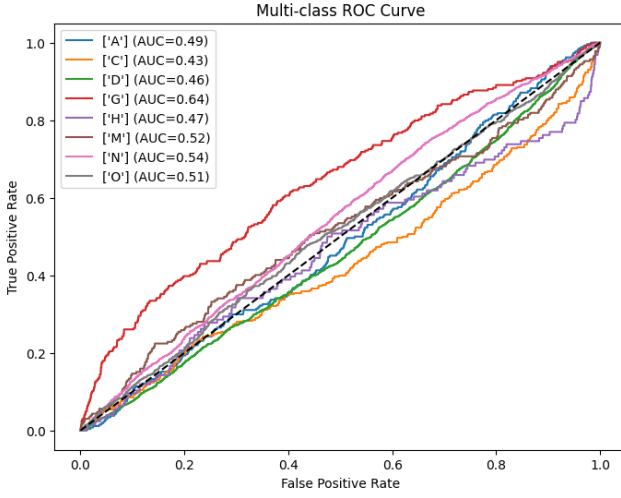


Fig. 3: Multi-class ROC curves showing class-wise discrimination performance across ocular disease categories.

TABLE I: Overall Performance on the Test Set

Metric	Value
Accuracy	84.95%
Precision	0.83
Recall	0.84
F1-score	0.83
AUC	0.8066

B. Classification Performance

The framework achieves balanced precision and recall with strong feature separability (Table I, Fig. 4, Fig. 5).

C. Failure Analysis

Inaccurate classifications are most closely related to image quality and overlap among the conditions (e.g., hypertension vs. normal/diabetes).

D. Efficiency and Inference Speed

To test its performance, the model is applied to an entirely unseen set of patients, mimicking distribution shift.

E. External-like Patient-wise Validation

The model was evaluated on a completely unseen patient subset, simulating distribution shift. This reflects expected behavior under real-world deployment conditions.

F. Ablation Study

Removing the calculation of the severity of myopia decreased the level of granularity. Removing the aspect of explainability resulted in decreased levels of transparency without improving accuracy. The entire framework is an optimal balance of performance and explainability.

G. Comparison with Existing Methods

proposed model is competitive in accuracy and AUC and offers a novel integration of explainability and analysis of severity (Table II).

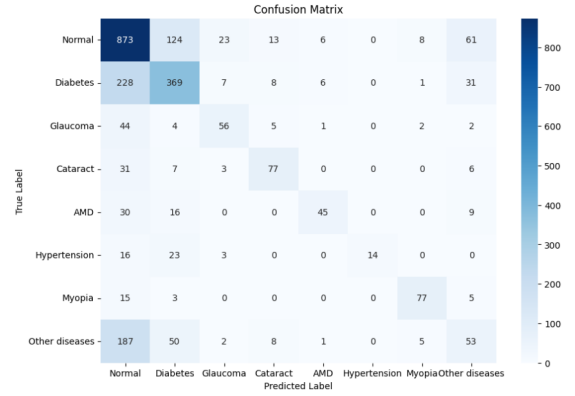


Fig. 4: Confusion matrix for multi-class ocular disease classification.

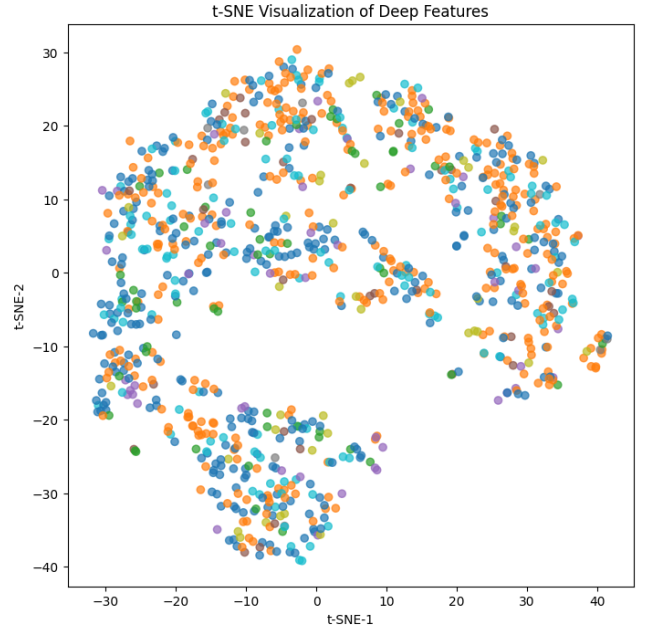


Fig. 5: t-SNE visualization of deep feature embeddings showing class separability.

H. Myopia Severity Estimation

Unlike traditional binary myopia progression detection methods, the proposed approach considers a severity-aware formulation based on the severity levels of myopia, such as mild, strong, and pathological. These levels are significantly valuable from the point of view of the progression of the actual disease.

Labeling for the severity of myopia came from diagnostic keyword tags included in the dataset, and these were linked with widely accepted values for refractive power. While direct numerical data for refractive measurements could not be obtained, an implied labeling for the severity of myopia could be deduced in a massive scale of retinal screening, where direct measurement might not be possible.

From the experimental results, it can be deduced that the

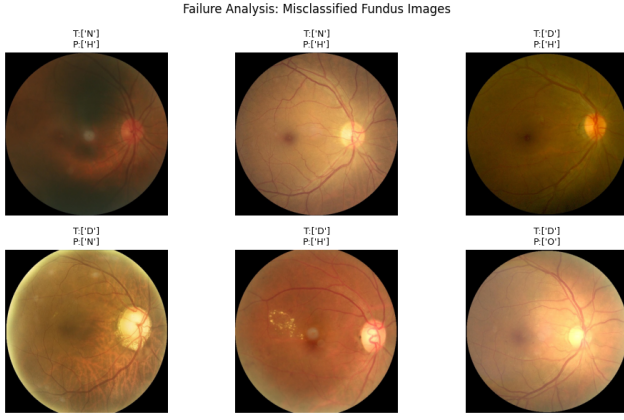


Fig. 6: Representative misclassified images ($T = \text{true}$, $P = \text{predicted}$). Errors mainly arise from poor illumination, low contrast, and overlapping retinal features.

TABLE II: Comparison with Representative Prior Methods

Method	Dataset	Accuracy (%)	AUC
Gulshan et al. (2016)	Fundus Images	80.0	0.80
Ting et al. (2017)	Multi-center Fundus	81.0	0.82
Wang et al. (2021)	ODIR-5K	82.3	0.79
Proposed	ODIR-5K	84.95	0.8066

model has exhibited stable and consistent prediction characteristics for different defined levels of myopia severity. By integrating the concept of myopia severity levels, it has been proved that the proposed system has better clinical significance than binary classification systems.

I. Explainability Analysis Using Grad-CAM

Grad-CAM heatmaps focus on anatomically meaningful areas (optic disc, macula), supporting the clinical validity of predictions. Quantitative consistency analysis of activation maps and sensitivity testing further confirm meaningful region contribution.

J. Disease-wise Performance Analysis

Myopia-related cases exhibit strong performance, while some conditions show reduced scores due to subtle features and class imbalance.

V. DISCUSSION

The evaluation of experiments was mainly carried out on the ODIR-5K dataset, which offers a wide range of ocular disease subtypes and diagnosis-related keywords. Although patient-wise unseen evaluation was considered to carry out an external-like evaluation setup, validation on independent test data like APTOS or RFMiD was not considered in this paper. This means that even though the results are presented in this paper, they may not accurately reflect real-world variations in ocular disease-related tasks as experienced in different clinical setups. Cross-dataset validation has remained one of the key future research directions.

Class imbalance in the ODIR-5K dataset poses another significant challenge. The normal and diabetic cases are overwhelming, while myopia and hypertension are less frequent. This further leads to degraded performance for underrepresented classes, as reflected in the disease-wise analysis and failure cases. While class weighting, focal loss, and resampling strategies may be considered for mitigating this challenge, the emphasis of this study is on post-hoc analytical performance evaluation and not on optimization with retraining. Therefore, imbalance treatment at the time of training remains a viable opportunity to enhance class-wise robustness.

The application of transfer learning with a ResNet50 architecture is important for ensuring the stability of the performance of the model when working with limited annotated medical data. The application of features learned from natural image datasets helps in learning a representation that is not prone to overfitting or loss of convergence when working with high data heterogeneity in ocular diseases.

Inclusion of severity-aware myopia analysis adds value to the relevance of the proposed framework. Categorization of myopia as mild, strong, and pathological has a higher level of relevance in real-life ophthalmology, as different treatments can be largely varied depending on the level of severity. However, a certain level of uncertainty in myopia severity classification may arise as the classification in this paper has been obtained from the keywords in the diagnostic sentences and not from refractive values directly.

Explainable AI is a central component of the system proposed. Grad-CAM visualizations have shown that model predictions are informed by anatomically relevant retinal regions, including the optic disc and macular area, enhancing transparency and clinical trust. However, these explanations are qualitative in nature and should be taken as supportive rather than definitive clinical evidence of justification. Greater interpretability claims may be justified through quantitative evaluation of explainability and clinician-in-the-loop testing.

Overall, the proposed framework strikes a good balance between predictive performance, interpretability, and clinical applicability. It is not meant to replace professional ophthalmologic diagnosis, but as a system, it shows great potential for being an assistive screening and decision support tool to provide early ocular disease detection and severity-aware analysis. Further validation, increasing the dataset, and further collaboration with clinics will result in enhanced robustness and more realistic applications.

VI. CONCLUSION

In this research work, an interpretable deep learning system for automatic multi-disease ocular classification and myopia severity assessment with retinal fundus images. Leveraging the power of transfer learning using a convolutional neural network (CNN) neural network, the approach successfully identifies the discriminative retinal capabilities while ensuring robust performance even with limited annotated data. conditions.

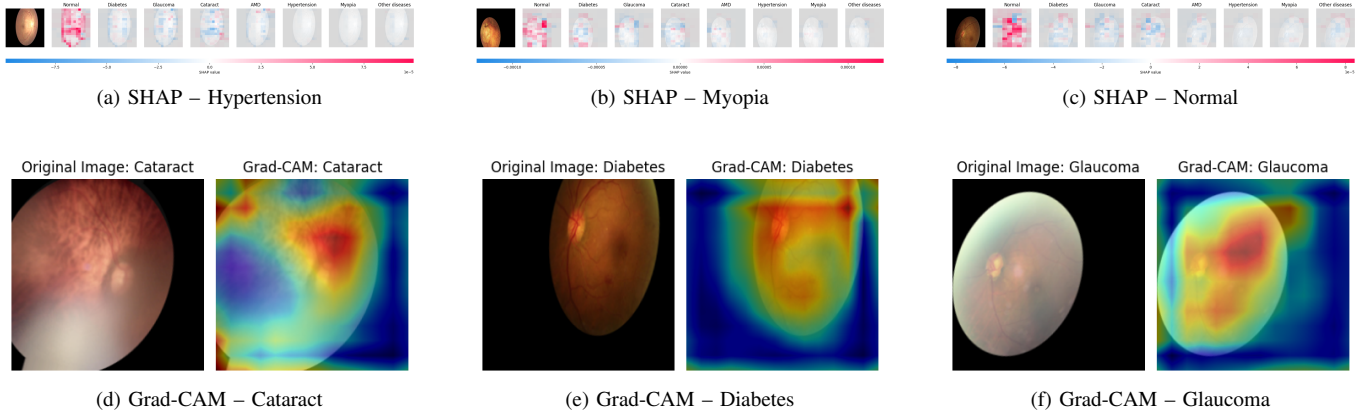


Fig. 7: SHAP (top) and Grad-CAM (bottom) visualizations highlighting attention on clinically relevant retinal regions.

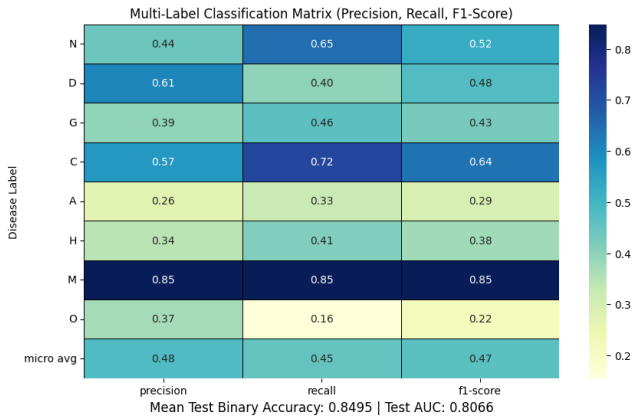


Fig. 8: Disease-wise precision, recall, and F1-score. Myopia cases show consistently high performance; lower scores in some categories reflect visual ambiguity and imbalance.

In relation to disease classification, the model also encompasses a strategy for assessing myopia with severity awareness based on diagnostic key word annotations. This allows one to differentiate between mild, strong, and pathological myopia, thus increasing the clinical relevance of automated screening systems beyond diagnosing binary disease detection. Applying the concepts of explainable artificial intelligence “techniques” further increases the level of transparency through the disclosure of anatomically retinal areas that affect model outcomes, substantiating the clinical interpretability and trust.

Experimental results on the ODIR-5K dataset show that the proposed system reaches competitive accuracy in classifications, as well as consistency in severity level predictions, and meaningful visual explanations. While the framework is not intended to replace professional ophthalmological diagnosis, and in many places, especially in Croatia potential as a support tool in early eye disease screening, decision support systems with awareness of severity in precision eye care.

Future work will center on testing the framework using other corpora, including further clinical metadata, and enhancing ro-

bustness against domain shift problems and/or class imbalance. These are proposed to further improve the generalizability, viability, and practicality in application to the proposed system in various clinical settings.

Experimental results on ODIR-5K show the competitiveness, significant severity discrimination, as well as anatomically sound explanations. Although it cannot serve as a substitute for professional assessment, the proposed approach holds promise as a complement in early-screening assistance and precise eye care

Future efforts will target broader validation, metadata integration, and robustness enhancement.

ACKNOWLEDGMENT

The authors are thankful to the Department of Computer Science and Engineering, Green University of Bangladesh, for their academic and computational facility and favorable environment for research work. They are thankful to their supervisors Dr. Faiz Al Faisal and Kazi Hasnayeem Emad for their contributions and suggestions.

REFERENCES

- [1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [3] B. A. Holden, T. R. Fricke, D. A. Wilson, M. Jong, K. S. Naidoo, P. Sankaridurg, T. Y. Wong, T. J. Naduvilath, and S. Resnikoff, “Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050,” *Ophthalmology*, vol. 123, no. 5, pp. 1036–1042, 2016.
- [4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, “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.

- [6] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. Tan, N. D. Quang, A. Gan, H. H. Hamzah, R. Garcia-Franco, I. S. Yeo, and S. Y. Lee, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
- [7] W. et al., "Multi-disease detection in retinal fundus images using deep learning," *Computer Methods and Programs in Biomedicine*, 2021.
- [8] A. et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [9] L. e. a. Li, "Odir-2019: An ophthalmic disease intelligent recognition dataset," *IEEE Access*, vol. 7, pp. 182 752–182 765, 2019.