RESEARCH PAPER

COMPARATIVE ANALYSIS OF DUAL-EYE CNN MODELS AND AN EXPLAINABLE MULTIMODAL CNN–VISION TRANSFORMER FOR OPHTHALMIC DISEASE DETECTION

Mr. Umesh Lakhtariya[1*], Dr. Ajay N. Upadhyaya[2]

[1,*]Research Scholar, Computer/IT Engineering Department, Gujarat Technological University, Ahmedabad, Gujarat, India.

[2]Professor, Computer Engineering Department, SAL Engineering & Technical Institute, SAL Education, Ahmedabad, Gujarat, India.

[1,*]https://orcid.org/0009-0000-4494-9443, [2]http://orcid.org/0000-0002-7583-6430

Email: [1,*]uplakhtariya46@gmail.com, [2]ajay8586g@gmail.com

ABSTRACT

Deep learning for the detection of ophthalmic disease has drawn much attention due to its promise for early diagnosis and automatic screening. Among existing approaches, dual-eye convolutional neural network (CNN) architectures have obtained promising performance by exploiting the bilateral ocular information from retinal images. However, the generalization capability and robustness of such models are still limited, especially across diverse datasets and disease categories. In this paper we report an extensive comparative study of dual-eye CNN-based models (DenseNet121, ResNet50) and traditional machine learning methods (Logistic Regression, Support Vector Machine, Random Forest, XGBoost) over several ophthalmic datasets. The experimental results show that the CNN-based models achieve higher accuracy and F1-score than the traditional methods, but there are still large performance gaps in recall and class-wise generalization, especially for complicated retinal conditions. To tackle these limitations, we propose an explainable multimodal hybrid framework that combines convolutional neural networks and Vision Transformers (ViT) with feature-level fusion. The proposed method aims to combine the ability of CNNs to extract spatial features with the ability of transformers to model global context, in addition to multimodal data including OCT, fundus images and clinical data. Moreover, explainable AI (XAI) techniques are expected to improve model interpretability and clinical trust. The results from our experimental analysis underline the importance of sophisticated hybrid architectures and strongly motivate the proposed framework. This work lays the groundwork for the development of more robust, interpretable and clinically applicable AI systems for ophthalmic disease detection.

Keywords: Ophthalmic Disease Detection, Convolutional Neural Networks (CNN), Vision Transformer (ViT), Multimodal Learning, Feature-Level Fusion, Explainable Artificial Intelligence (XAI), Retinal Image Analysis.

How to cite this article: Lakhtariya U, Upadhyaya AN. Comparative Analysis of Dual-Eye CNN Models and an Explainable Multimodal CNN–Vision Transformer for Ophthalmic Disease Detection. Int J Drug Deliv Technol. 2026;16(53s): 350-358. DOI: 10.25258/ijddt.16.53s.37

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

Diabetic retinopathy (DR) and Glaucoma are some of the most common ophthalmic diseases in the world, leading to irreversible vision loss. Early and accurate diagnosis is crucial to halt the disease progression. In the last few years, deep learning techniques, particularly convolutional neural networks (CNNs), have been successfully applied to the automated analysis of retinal images, owing to their ability to learn effective feature extraction from fundus and optical coherence tomography (OCT) images. These advances have been driven by publicly available benchmark datasets such as the APTOS 2019 Blindness Detection Dataset [1] and the ACRIMA Glaucoma Dataset [2] that are commonly used for diabetic retinopathy and glaucoma detection tasks. The CNN-based approaches have substantially improved the performance of ophthalmic disease classification systems and are widely adopted because of their ability to learn hierarchical feature representations from medical images [3], [4]. However, the traditional single input CNN models usually cannot capture the complementary diagnostic information in bilateral retinal images. To overcome this limitation, dual-eye CNN architectures have been proposed to process the images of both eyes together, which enhances the classification performance. The models exploit correlations between the two eyes and have been shown to perform better than models using only one eye. Their performance is still limited due to poor generalization on different datasets and the differences in disease distribution [5], [6]. In addition, CNN-based ophthalmic models often suffer from reduced recall for minority classes and performance degradation when evaluated on unseen clinical data, which poses challenges for real-world deployment [7].

Besides these limitations, CNN-based models primarily extract local spatial features, which restricts their capacity to capture the global contextual relationships necessary for identifying complex retinal patterns. Vision Transformers (ViTs) based on self-attention mechanisms have emerged as a strong alternative, capable of modeling long-range dependencies in image data. The effectiveness of transformer-based architectures for medical image analysis, including ophthalmic imaging tasks, has been shown by recent studies [8], [9]. However, pure ViT models usually need large-scale datasets and may not be efficient to capture fine-grained local features. To tackle these challenges, recent studies investigated hybrid architectures that combine CNNs and transformers to benefit from the advantages of both local feature extraction and global context modelling [10]. Although such approaches have shown promising improvements, their application in ophthalmic disease detection, especially in

multimodal settings including fundus images, OCT scans, and clinical data, is limited and under-explored.

Another major challenge with the current ophthalmic AI systems is lack of interpretability. Transparency and trust are required in clinical practice, but most deep learning models are black-box systems that hinder their adoption. Explainable Artificial Intelligence (XAI) techniques have been proposed to provide visual and feature-level explanations, but their integration into hybrid multimodal frameworks is still in its infancy [11]. Moreover, most of the existing studies are based on single modality inputs such as fundus images, and do not exploit complementary diagnostic information such as OCT scans or patient clinical data. This restricts the model from fully capturing the complexity of ophthalmic diseases, which are inherently multimodal. In addition, the ineffective feature-level fusion strategies lead to poor performance in complicated diagnostic scenarios [12].

To overcome these limitations, this study provides a comparative study of dual-eye CNN models, namely DenseNet121 and ResNet50, and traditional machine learning classifiers, namely Logistic Regression, Support Vector Machine, Random Forest and XGBoost, on benchmark datasets, namely the APTOS 2019 Blindness Detection Dataset [1] and the ACRIMA Glaucoma Dataset [2]. The experimental results show large performance gaps on recall, generalization, and handling of class imbalance. In view of these observations, this work proposes an explainable multimodal hybrid framework with feature-level fusion by combining CNN and Vision Transformer (ViT) architectures. The proposed framework aims at combining fundus images, OCT scans and clinical attributes to improve diagnostic accuracy and applying XAI techniques for improving interpretability and clinical trust. Thus, the study provides a critical evaluation of current approaches and a foundation for the development of more robust and clinically applicable ophthalmic AI systems.

I.1 CONTRIBUTIONS OF THE STUDY

This study conducts a systematic study on deep learning-based ophthalmic disease detection and proposes an advanced hybrid framework to overcome the existing limitations. The main contributions of this work are summarized below:

1. **Fully Deploying Dual Eye CNN Baselines:** We describe the implementation and assessment of dual eye inspired convolutional neural network architectures based on DenseNet121 and ResNet50 for ophthalmic disease detection. We reproduce baseline methods and adapt them to guarantee experimental reliability and reproducibility on benchmark datasets.
2. **Comparison with Traditional Machine Learning Models:** Apart from deep learning models, traditional machine learning classifiers such as Logistic Regression, Support Vector Machine, Random Forest and XGBoost are also used with deep feature embeddings. The performance differences between the methodologies are illustrated by a detailed comparative analysis.
3. **Extensive experimental assessment on benchmark datasets:** The models are thoroughly assessed on publicly available datasets, i.e., APTOS 2019 Blindness Detection Dataset and ACRIMA Glaucoma Dataset in multi-class (diabetic retinopathy grading) and binary (glaucoma detection) classification tasks. We test the performance with standard metrics like accuracy, precision, recall and f1-score.
4. **Identified Performance Gaps:** The experimental analysis in this study reveals significant shortcomings of existing approaches, such as low recall for minority classes, misclassification between neighboring disease severity levels, and poor generalization across datasets. The analysis of the confusion matrix and metric trends further support these findings.
5. **Analysis of Limitations of CNNs in Modeling Global Context:** The study found that standalone CNN models, while showing significant improvement over traditional methods, are limited in their ability to model long-range dependencies and global retinal structures that are important for precise ophthalmic diagnosis.
6. **Motivation of Hybrid CNN-Vision Transformer Architectures:** The shortcomings observed here give a strong motivation for the combination of the Vision Transformers (ViTs) with CNNs to exploit the local feature extraction property and the global contextual reasoning power.
7. **Proposed Explainable Multimodal Hybrid Framework:** We propose a novel hybrid framework that combines the CNN and Vision Transformer architectures with feature-level fusion. The framework is designed to incorporate multi-modal inputs, including fundus images, OCT scans and clinical attributes, to improve the diagnostic performance.
8. **Integration of Explainable Artificial Intelligence (XAI):** The proposed framework integrates explainability mechanisms to improve transparency and interpretability, thus increasing clinical trust and adoption in real-world healthcare settings.
9. **Towards Future Multimodal Ophthalmic AI Systems:** The study proposes a systematic pathway from baseline implementation to advanced hybrid modeling, laying the foundation for future work on interpretable and multimodal ophthalmic disease detection.

II. LITERATURE REVIEW

The rapid development of deep learning has transformed the field of ophthalmic disease detection through the analysis of retinal images, including fundus photographs and OCT scans. Convolutional neural networks (CNNs) have been widely utilized because of their powerful hierarchical feature extraction capability, which makes them suitable for accurate disease classification such as diabetic retinopathy and glaucoma. CNN-based models have shown a high accuracy for the retinal image classification tasks, and hence are suitable for automated screening systems [3], [4]. However, some limitations of CNN based approaches have been questioned in recent literature. These models mainly rely on local spatial features and often ignore modeling the global contextual relationships in retinal images. This limitation is more severe in the case of the multi-class classification problems such as diabetic retinopathy grading where subtle differences between adjacent classes lead to misclassification [5], [6]. Furthermore, CNN-based models are also plagued by the problem of class imbalance and low recall of minority classes, which limits their clinical application [7].

To overcome these challenges, researchers have investigated transformer-based architectures, especially Vision Transformers (ViTs), which utilize self-attention mechanisms to capture long-range dependencies in image data. ViTs have demonstrated their promising performance on various medical imaging tasks by better capturing global contextual information than traditional CNNs [8], [9]. However, vanilla transformer models require large datasets, and are not effective at capturing fine-grained local features that are important in ophthalmic image analysis. Recent work has proposed hybrid architectures combining CNNs and transformers to take advantage of the benefits of both approaches. These models leverage the local feature extraction ability of CNNs and the global reasoning ability of the transformers, achieving better performance for complex image classification tasks [10]. However, the use of such hybrid models in ophthalmology is still limited, especially in the case of multimodal data integration.

Another trend in medical AI research is multimodal learning, where multiple data sources like fundus images, OCT scans, and clinical attributes are integrated to enhance diagnostic accuracy. Multimodal approaches have shown to be able to capture complementary information across the modalities, leading to more robust predictions [12]. However, the effective feature-level fusion strategies and their integration with advanced deep learning architectures are still underexplored in the ophthalmic applications. In addition to performance considerations, the interpretability of deep learning models remains a significant issue for clinical applications. Explainable Artificial Intelligence (XAI) techniques have been proposed to address this problem by providing visual explanations and feature importance information. While XAI applied to standalone CNN or transformer models has been shown to be successful, the combination of XAI with hybrid multimodal frameworks is still in its infancy [11].

In general, the review highlights the significance of CNNs and transformer-based models for ophthalmic disease detection, however, there are major gaps in modeling global context, multimodal data, and interpretability. Such limitations highlight the need for a unified framework that combines CNNs, Vision Transformers, multimodal learning and explainable AI to improve diagnostic performance and clinical reliability.

Table 2.1. Comparative Analysis of Existing Approaches

Ref	Methodology	Data Type	Strengths	Limitations
[3]	CNN-based deep learning	Fundus images	High accuracy in DR detection	Limited interpretability
[4]	CNN for glaucoma detection	Fundus images	Effective optic disc analysis	Poor generalization
[5]	Deep learning in healthcare	Multi-domain	Strong feature learning	Requires large datasets
[6]	Multi-view CNN	Retinal images	Captures multiple perspectives	Limited global context
[8]	Transformer-based models	Image data	Captures global dependencies	High data requirement
[7]	Vision Transformer (ViT)	Image classification	Strong global attention	Weak local feature capture
[10]	CNN + Transformer hybrid	Medical images	Improved performance	Limited ophthalmic application
[11]	Explainable AI	Healthcare data	Enhances interpretability	Not integrated with multimodal DL
[12]	Multimodal deep learning	Imaging + clinical	Improved robustness	Fusion complexity
[13]	Swin Transformer for medical imaging	Fundus + OCT	Strong hierarchical attention	Requires large data
[14]	CNN–ViT hybrid architecture	Retinal images	Combines local + global features	Computational complexity
[15]	Multimodal fusion network	Fundus + clinical	Improved diagnostic accuracy	Fusion design challenges
[16]	Attention-based multimodal DL	Imaging + EHR	Better feature interaction	Limited explainability
[17]	Explainable CNN with Grad-CAM	Fundus images	Visual interpretability	Limited global reasoning
[18]	Transformer-based DR grading	Fundus images	Improved multi-class performance	Data intensive
[19]	Hybrid DL with feature fusion	Retinal imaging	Robust classification	Lack of multimodal integration
[20]	XAI-integrated transformer	Medical imaging	Enhanced interpretability	Computational overhead

The comparison lucidly shows that recent research is gradually moving towards hybrid CNN-Transformer architectures and multimodal learning frameworks to address the limitations of traditional methods. Transformer-based models have demonstrated their power in modeling global dependencies . Multimodal systems can enhance the robustness of diagnosis by fusing complementary data sources . However, these approaches still face challenges of computational complexity, effective feature fusion, and limited integration of explainability mechanisms. More interestingly, only a few studies jointly consider the three crucial aspects of global context modeling, multimodal data fusion, and explainability in a unified framework. This gap means that a comprehensive hybrid architecture of CNNs, Vision Transformers, multimodal inputs and

explainable AI is needed which is the basis of the approach put forward this study.

II.1 RESEARCH GAP IDENTIFIED

The above literature review shows the following research gaps:

1. **Absence of Global Context Modeling in CNN-based Ophthalmic Systems:** The current CNN models are unable to model long-range spatial dependencies that are necessary to accurately grade diseases.
2. **Underutilization of Hybrid CNN-Transformer Models in Ophthalmology:** Hybrid models have potential but have not been widely explored for the detection of retinal diseases.
3. **Unused Multimodal Data:** Most of the studies only use a single type of input, discarding the additional information that can be obtained from OCT and clinical data.
4. **Explainability in Multimodal Hybrid Architectures:** Current models lack embedded XAI mechanisms, which limits their clinical trustworthiness.
5. **Generalization and class-imbalance:** Existing methods face challenges to model the diversity of the real world and to predict the minority classes.

III. METHODOLOGY

In this section, the methodology of this work is described to evaluate the current deep learning techniques for ophthalmic disease detection, and to propose a more advanced hybrid solution. The proposed method systematically evaluates the performance of traditional machine learning models and convolutional neural networks (CNNs) on benchmark retinal datasets. A detailed analysis of the limitations of these approaches in terms of generalization, handling class imbalance and representation of contextual features is provided. Therefore, we propose a hybrid

multimodal architecture of CNNs and Vision Transformers (ViTs) with explainable artificial intelligence (XAI). This strategy achieves a reasonable trade-off between empirical evaluation and methodological advancement and offers a comprehensive pathway from baseline implementation to the development of a more robust and interpretable ophthalmic diagnostic framework.

III.1 DATASET DESCRIPTION AND PREPROCESSING

In this study, we evaluate ophthalmic disease detection models using two publicly available benchmark datasets: the APTOS 2019 Blindness Detection Dataset [1], which contains fundus images labeled across five diabetic retinopathy severity levels (0–4), and the ACRIMA Glaucoma Dataset [2], which consists of retinal images categorized into normal and glaucomatous classes. These datasets allow for the assessment of
both multi-class and binary classification tasks, offering a complete testing environment for model performance at different complexity levels.

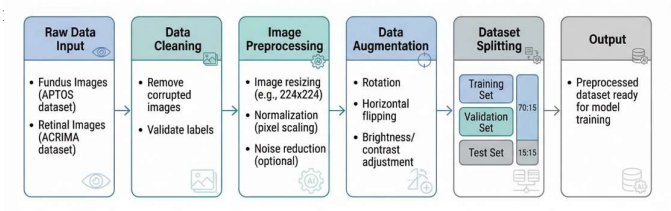

Figure 1: Data Preparation and Preprocessing Pipeline

Before training the model, all images are preprocessed in a standardized way to maintain consistency across datasets. The images are resized to a fixed resolution and normalized to aid convergence during training. Data cleaning procedures are applied to eliminate corrupted or invalid samples, followed by stratified splitting into training, validation, and test sets to preserve class distribution. Data augmentation techniques like rotation, horizontal flipping, and intensity variations are applied to improve generalization due to the inherent class imbalance in diabetic retinopathy grading. The complete preprocessing and data pipeline used in this study is shown in Figure 1.

III.2 BASELINE MODEL DEVELOPMENT AND EXPERIMENTAL CONFIGURATION

Both deep learning and traditional machine learning approaches are used to create a robust experimental baseline. Pretrained convolutional neural networks (ResNet50 and DenseNet121) are tuned on the target datasets by replacing the last fully connected layers to match the respective classification task. These models are trained with the cross-entropy loss function, defined as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i is the true label and y_i is the predicted probability for class i. The optimization process is carried out with the Adam optimizer with a learning rate that is controlled to ensure convergence. We extract feature embeddings from CNN backbones to train classical classifiers such as Logistic Regression, Support Vector Machine, Random Forest and XGBoost as well as deep learning models. This enables a fair comparison of shallow learning methods and deep neural networks with consistent feature representations.

The experimental setup is done with a batch size of 32 and it is trained for a limited number of epochs to avoid overfitting but also to learn enough. We evaluate the model performance with standard classification metrics including accuracy, precision, recall and F1-score. The metrics are calculated as follows:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}, \quad F1 = \frac{2 \cdot (Precision \cdot Recall)}{Precision+Recal}$$

The comparative analysis of traditional machine learning models and convolutional neural networks (CNNs) was performed to evaluate the effectiveness of different learning paradigms for both classification tasks. The models were assessed based on important performance metrics, i.e., accuracy and F1-score, to quantify the overall correctness and the balance among classes,

respectively. This comparison is particularly important in the case of ophthalmic disease detection, where class imbalance and small inter-class variations can have a significant impact on the reliability of models. The results of this comparison are shown in Figure 2 and reveal the performance differences across models and tasks.

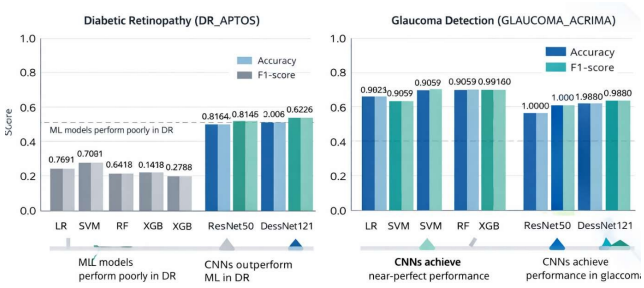

Figure 2: comparative performance of ml and cnn models (accuracy and f1-score)

As shown in Figure 2, CNN based models outperform the traditional machine learning methods in both tasks. The improvement is more significant in the multi-class classification case. Traditional models have moderate accuracy for diabetic retinopathy, but their F1-scores are low, indicating poor class-wise discrimination. CNN models demonstrate a substantial improvement in both metrics. By contrast, all models perform well on the binary classification task, with CNN architectures performing near-perfectly. This difference represents the effect of task difficulty and also reflects the insufficient capacity of traditional methods and CNN-based methods in capturing complex feature relations, which motivates the study on more advanced hybrid architectures.

III.3 PROPOSED HYBRID MULTIMODAL CNN–VISION TRANSFORMER FRAMEWORK

The limitations revealed in the baseline experimental analysis, particularly in handling complicated multi-class classification and global contextual dependencies, motivate the design of a hybrid multimodal architecture integrating convolutional neural networks (CNNs), Vision Transformers (ViTs) and explainable artificial intelligence (XAI) methods. The proposed framework aims at exploiting the advantages of different learning paradigms and heterogenous data sources to improve the diagnostic accuracy, robustness and interpretability. The architecture begins with a multi-modal input layer that combines various sources of ophthalmic data, such as fundus images, optical coherence tomography (OCT) scans, and structured clinical attributes. Each modality provides different but complementary information. Fundus images provide surface level information of the retina, OCT scans provide cross-sectional structural information and the clinical data provides contextual information about the patient's condition and history. The modalities are integrated to address the limitations of single modality systems, and to provide a more complete characterization of disease.

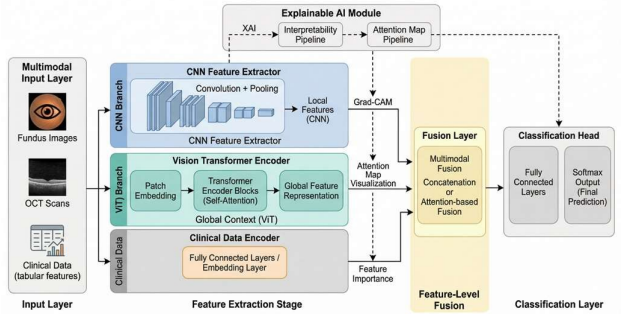

Figure 3: Proposed Hybrid Multimodal CNN–ViT Architecture with Explainable AI Integration

In the feature extraction phase, a CNN-based branch is employed to process image-based inputs, especially fundus images, to extract fine-grained local spatial features like lesions, microaneurysms and exudates. The convolutional and pooling operations in CNNs effectively learn the hierarchical feature representations, which help the model to recognize the important visual patterns related to disease progression. However, as shown in the experimental results, CNNs may not be good at capturing long-range dependencies and global contextual relationships across the retinal image.

To overcome this limitation, the proposed framework consists of a Vision Transformer (ViT) branch which processes the image data by modeling relationships between different image regions with self-attention mechanisms. The transformer architecture, in contrast to CNNs, allows the model to learn global dependencies and spatial interactions across the entire image, which is critical to identify disease severity patterns that are distributed over several regions. The combination of CNN-ViT ensures the effective capture of local and global features. Meanwhile, clinical data are processed by a dedicated encoding module based on fully connected layers to convert the structured inputs into a dense feature representation. This enables the model to include non-imaging information such as patient-specific attributes that can greatly enhance the diagnostic performance when combined with visual features.

Then, features extracted from all modalities are fused through a feature-level fusion mechanism. In the fusion step, the outputs of the CNN, ViT, and clinical data encoder are combined into one representation, which allows the model to utilize the complementary information from the different sources. The fusion layer is important to improve the discrimination ability of the model, especially in complicated classification tasks where the single-modality features are insufficient. The joint feature representation after fusion is then fed into a classification head consisting of fully connected layers to generate the final prediction. This stage maps the learned features to the class probabilities for different disease categories.

The proposed framework employs explainable AI techniques at multiple levels to promote transparency and clinical trust. The CNN branch is fed into grad-CAM to highlight important regions in fundus images, which can be visually interpreted by clinicians to diagnose the model. Similarly, attention maps from the transformer allow us to gain insight into global feature interaction and feature importance analysis is used to interpret the contribution of clinical variables. The multi-level explainability increases the interpretability of the system and tackles one of the main limitations of current black-box deep learning models.

In summary, the proposed hybrid multimodal framework integrates local feature extraction, global context modeling, and multimodal data integration in a single architecture, and also includes explainability to support clinical decision making. This approach directly addresses the limitations seen in the baseline models, thereby providing a solid foundation towards the development of more accurate, robust, and explainable AI-based systems for ophthalmic disease detection.

IV. RESULTS AND DISCUSSIONS

IV.1 EXPERIMENTAL RESULTS

The experimental evaluation is performed on two benchmark datasets [1] APTOS 2019 Blindness Detection Dataset for diabetic retinopathy and [2] ACRIMA Glaucoma Dataset for glaucoma detection. Standard evaluation metrics such as accuracy, precision, recall and F1-score are used to evaluate both traditional machine learning models and deep learning architectures.

Traditional machine learning models have been used to classify diabetic retinopathy and achieved moderate performance, with Logistic Regression having the highest F1-score of about 0.50, whereas other models such as Support Vector Machine, Random Forest and XGBoost have lower recall and F1-scores. This indicates the inability of shallow classifiers to encode complex retinal patterns although they are trained on deep feature embeddings. On the other hand, the CNN-based models to significantly improve the performance. ResNet50 and DenseNet121 achieve the test accuracy of ~0.816 and 0.814, and the F1-score of 0.6108 and 0.6226, respectively. The results demonstrate the effectiveness of deep feature learning in solving complex multi-class classification problems.

For glaucoma detection, both the traditional machine learning and CNN models show much higher performance than diabetic retinopathy classification. XGBoost achieves F1-score around 0.916 whereas CNN models demonstrate near-perfect classification with ResNet50 achieving F1-score of 1.000 and DenseNet121 achieving 0.988. This performance can be explained by the binary nature of the task and the existence of strong structural features such as optic disc and cup-to-disc ratio that are more easily identified than subtle lesion patterns in diabetic retinopathy.

To provide additional insights into the classification behavior of the models beyond aggregate metrics, confusion matrices are examined for both multi-class and binary classification tasks. However, confusion matrices give a more detailed class-wise understanding of the prediction performance including the correct classifications and the misclassification patterns. This analysis is particularly important in detecting ophthalmic diseases where it is often difficult to distinguish between adjacent levels of disease severity. Figure 4 shows the confusion matrices for both tasks, which permit a more detailed analysis of the model reliability and error distribution.

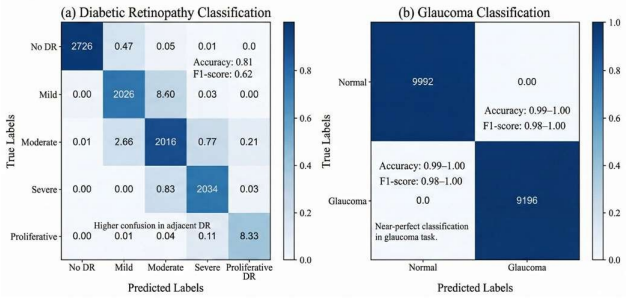

Figure 4: Confusion Matrix for Diabetic Retinopathy and Glaucoma Classification

As shown in Figure 4, in the multi-class classification task, there is a significant misclassification between adjacent severity levels, particularly between mild, moderate, and severe classes. The diagonal values show that the model can correctly identify most of the instances, while the off-diagonal values show that the model struggles to capture fine-grained distinctions between similar classes. This is consistent with the moderate F1-score obtained for this task and highlights the inherent complexity of multi-class disease grading.

In comparison, the binary classification task yields almost perfect results: almost all samples are classified correctly and there are almost no misclassifications. Strong diagonal dominance in the confusion matrix is an indication of high model confidence and clear separability between the two classes. However, such high performance can also be affected by the simplicity of the task and dataset characteristics, and does not guarantee robustness in more complex or real-world scenarios. These results emphasize the need for more sophisticated architectures that can capture complex feature relations and improve class-wise discrimination in challenging classification scenarios.

To comprehensively evaluate the model performance, we performed a comparative analysis on multiple evaluation metrics, namely accuracy, precision, recall and F1-score. Single metrics provide a focused piece of information, but visualizing them together allows for a comprehensive view of model performance on different tasks and learning modes. In particular it shows the trade-off between precision and recall and the overall trade-off measured by the F1-score. Figure 5 shows the comparison of the performance of all the models in both classification tasks. It is worth noting that the figure highlights the important trends in the performance of models and not the specific numeric values.

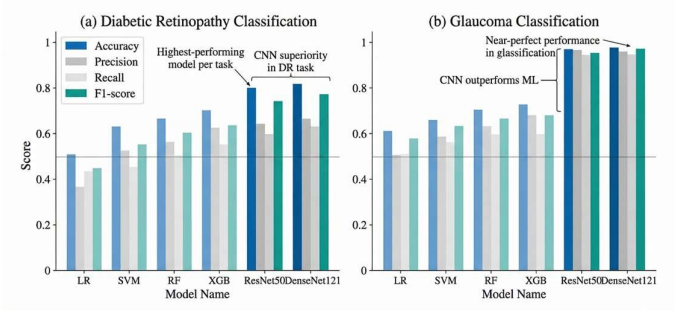

Figure 5: Performance Comparison of ML and CNN Models (Accuracy, Precision, Recall, F1-Score)

It can be seen from Figure 5 that CNN-based models outperform traditional machine learning methods on all evaluation metrics, and the gap is even larger for the multi-class classification task. Here we observe that conventional models have lower recall and F1-scores, indicating that it is difficult to make balanced predictions for each class, and CNN models show a more stable and consistent performance over different metrics. On the other hand, the binary classification task is well handled by all models, with little variation across metrics, indicating the relatively easy nature of the problem. Such observations reinforce the limitations of existing approaches in handling complex classification scenarios, and further motivate the requirement of advanced hybrid architectures to improve both feature representation and class-wise discrimination.

IV.2 DISCUSSION AND ANALYSIS

Experimental results show a clear and consistent distinction in the performance of models on the two classification tasks, providing insights into the strengths and limitations of traditional machine learning approaches as well as deep learning architectures. As shown in Figure 2 and Figure 5, CNN-based models have a significant improvement over conventional methods in all evaluation metrics, validating their ability in extracting complex visual patterns from retinal images. But despite this improvement, the performance achieved on the diabetic retinopathy classification task is still moderate, especially in terms of recall and F1-score. This means that while CNNs are capable of capturing discriminative features, they are still limited to obtain balanced classification across all disease categories. This is due to the multi-class nature of the problem, subtle inter-class variations and class imbalance, which reduces the overall robustness of the model.

Figure 4 shows the confusion matrix that gives a deeper look into the classification behaviour and further reinforces these limitations. The results clearly show that misclassification is common between adjacent severity levels like mild and moderate, or moderate and severe stages. This pattern indicates that the model has difficulty with the fine-grained difference of lesion characteristics and their spatial distribution over the retina. The diagonal dominance of the confusion matrix shows that the model is performing reasonably well overall, but the significant off-diagonal values show that the model has not learned the exact boundaries between closely related classes. This problem is especially acute in clinical applications, where correct grading of disease severity is essential for treatment planning and prognosis.

In contrast, the glaucoma classification task exhibits near-perfect performance of all models, where CNN-based approaches achieve almost complete accuracy and F1-scores, as shown in Figures 2 and 5. This is likely attributed to the binary nature of the task and the presence of easily-distinguishable structural features such as optic disc and cup-to-disc ratio, as opposed to more subtle and distributed patterns present in diabetic retinopathy. This is also supported by the confusion matrix of the task which shows strong diagonal dominance and negligible misclassification. These results show that the model is highly reliable on the dataset used, however they should be interpreted with caution. Such near-perfect performance on curated datasets may not necessarily translate to real-world clinical environments where variations in image quality, patient demographics and acquisition conditions can significantly impact model performance. Thus, the generalization ability of these models is still a matter of concern.

One key limitation underlying the performance gap is the intrinsic architecture limitation of CNN-based models that mainly extract features via local receptive fields. This local processing is useful for detecting small scale patterns such as lesions and textures but limits the model to capture long-range dependencies and global contextual relationships across the whole retinal image. This limitation is especially significant in complex classification problems, such as grading diabetic retinopathy, where the severity of the disease is dependent not only on the existence of local abnormalities, but also on the spatial distribution and interaction of the abnormalities over different regions of the retina. These global relationships are not well modeled, resulting in lower discriminative capability, and partially explaining the misclassification patterns observed in the results.

Moreover, the difference in the performance of classes reflects the effect of data imbalance and poor representation of minority classes. The recall and F1-score of traditional Machine Learning models drastically drop which indicates poor generalization and sensitivity to under-represented categories. CNN models partially solve this problem by learning hierarchical features, but they are still not able to perform consistently across all classes. This imbalance affects not only the accuracy of classification but also the clinical reliability of the system, as the minority classes usually represent critical stages of disease that need timely intervention. Apart from the architectural and data-related challenges, the interpretability of deep learning models is another critical challenge. CNNs give better performance but they are black-box models and it is difficult to understand the reasoning behind the predictions especially when there is a misclassification. This opacity could be a problem for clinical adoption, as medical practitioners need clear and explainable evidence to trust automated diagnostic systems. The absence of interpretability mechanisms within the baseline models underlines the significance of incorporating explainable AI techniques into the diagnostic framework.

To sum up, the detailed analysis of results in several figures clearly shows that CNN-based models outperform traditional methods but are not enough to solve the complex

problems of detecting ophthalmic diseases in real world scenarios. Existing methodologies have limitations in capturing global context, handling class imbalance and providing interpretability which highlights critical gaps. These results strongly motivate the development of more sophisticated architectures that can combine local and global feature learning, use multimodal data sources and provide clear decision making processes. These challenges motivate the proposed hybrid multimodal CNN-Vision Transformer framework with explainable AI integration for better performance and clinical applicability.

IV.3 IMPLICATIONS FOR HYBRID CNN–VISION TRANSFORMER FRAMEWORK

The experimental results provide strong evidence for more sophisticated architectures beyond the conventional CNN-based methods . The results show that although CNN models are better than traditional machine learning methods, they still have difficulty in complex multi-class classification tasks, especially in cases that require fine-grained distinction and balanced class-wise performance. The main reason of these limitations is the local receptive field of CNNs, which limits the ability of CNNs to learn long-range dependencies and global contextual relationships in retinal images. The use of Vision Transformers (ViTs) here is an encouraging direction to address these challenges. Unlike CNNs, transformer-based architectures are designed to model global interactions with self-attention mechanisms, and thus can better understand spatial dependencies across the whole image. Thus, it is anticipated that the integration of CNNs and ViTs will provide a complementary learning framework, where CNNs are responsible for extracting detailed local features and ViTs are responsible for extracting global contextual information, resulting in enhanced feature representation and classification performance.

The experimental results also show the limitation of using single-modality data alone, especially for complex diagnostic tasks. Incorporating multimodal inputs, such as OCT scans and clinical attributes, can offer supplementary context that enhances the model's capacity to distinguish between similar disease stages. Such multi-modal integration is expected to make it more robust, reduce the misclassification and support more accurate clinical decision making. Another major implication of the analysis is the need for better interpretability of models. Deep learning models perform well but the black-box nature is a major barrier to clinical adoption. The proposed framework incorporates explainable artificial intelligence (XAI) methods to provide transparency by providing visual and feature-level explanations of model predictions. This builds trust and also helps in validating the .

The experimental analysis generally confirms the efficiency of CNN based models, but also clearly shows their limitations in capturing global context, handling class imbalance and interpretability. The results justify the proposed hybrid multimodal CNN-Vision Transformer framework with explainable AI integration to address these challenges and promote the development of reliable and clinically applicable ophthalmic diagnostic systems.

IV.4 COMPARATIVE INSIGHTS

The comparative evaluation over several figures depicts a consistent and interpretable trend in performance, emphasizing the superiority of deep learning methods over traditional machine learning models in ophthalmic disease detection. From Figure 2 and Figure 5, it can be seen that the accuracy and F1-scores are higher and more balanced for CNN-based models, especially for the multi-class classification task, which indicates that CNN-based models are better in learning complex visual patterns. However, the confusion matrix in Fig. 4 indicates that even these models have difficulty in fine-grained discrimination between adjacent disease stages, which is an unavoidable complexity of the task. In contrast, the binary classification task exhibits a high performance for all models. This suggests that the classification task is considerably simpler when the decision boundaries are simpler and the features are well defined. Overall, these insights highlight that CNNs are an important step forward, but not sufficient in themselves to solve complex clinical scenarios; this further underlines the need for hybrid architectures that combine global context modeling, multimodal information and explainability to achieve more robust and clinically reliable results.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we present a detailed review of deep learning techniques for detecting ophthalmic diseases, emphasizing traditional machine learning models and convolutional neural network (CNN) architectures. The experimental results show that the CNN-based models achieve much better overall classification performance than the traditional methods. However, the results also point to critical limitations, particularly in complex multi-class classification settings, where moderate performance and frequent misclassification between closely related disease stages continue to be observed. The analysis indicates that although CNNs are good at capturing local spatial features, they are fundamentally limited in modeling global contextual relationships in retinal images. This, in addition to the problems of class imbalance and generalization, limits their application to real-world clinical applications. On the other hand, simpler classification problems achieve a much higher performance, demonstrating the impact of task complexity and data characteristics on the model performance. These observations highlight the need for more advanced architectures to learn the local and global feature dependencies.

Motivated by these findings, this work proposes a hybrid multimodal framework that combines CNNs with Vision Transformers (ViTs) for combining local feature extraction and global context modeling. The proposed framework is also designed to incorporate multiple data modalities and explainable artificial intelligence techniques to improve interpretability and clinical reliability. Even though the proposed model is not tested in this study, it offers a clear and well-reasoned direction to overcome the identified limitations. Future work will involve the implementation and validation of the proposed hybrid framework, including the integration of multimodal data sources and sophisticated feature fusion strategies. Further research will also explore the robustness and generalizability of the model across different clinical settings. Furthermore, the incorporation of explainability mechanisms will be explored for the sake of transparency and supporting clinical decision making. In summary, this work provides a foundation for the development of more accurate, reliable and interpretable AI-based systems for detection of ophthalmic diseases.

VI. REFERENCES

[1] Kaggle. (2019). *APTOS 2019 blindness detection dataset*. https://www.kaggle.com/datasets/valmmm/aptos2019

[2] Kaggle. (2020). *ACRIMA glaucoma dataset*. https://www.kaggle.com/datasets/sohamchakraborty2004/acrima-glaucoma

[3] Jin, K., & Ye, J. (2022). Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives. Advances in Ophthalmology Practice and Research, 2(3), 100078. https://doi.org/10.1016/j.aopr.2022.100078

[4] Li, Z., He, Y., Keel, S., Meng, W., Chang, R. T., & He, M. (2018). Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color Fundus photographs. Ophthalmology, 125(8), 1199–1206. https://doi.org/10.1016/j.ophtha.2018.01.023

[5] Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable Deep Learning: a field guide for the uninitiated. Journal of Artificial Intelligence Research, 73, 329–397. https://doi.org/10.1613/jair.1.13200

[6] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., & Qu, R. (2019). A survey of Deep Learning-Based Object Detection. IEEE Access, 7, 128837–128868. https://doi.org/10.1109/access.2019.2939201

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (n.d.). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. OpenReview. https://openreview.net/forum?id=YicbFdNTTy

[8] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022), 41 pages. https://doi.org/10.1145/3505244

[9] Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020, April 24). Lite Transformer with Long-Short Range Attention. arXiv.org. https://arxiv.org/abs/2004.11886

[10] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNET: Transformers make strong encoders for medical image segmentation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.04306

[11] Cheng, Z., Wu, Y., Li, Y., Cai, L., & Ihnaini, B. (2025). A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. Sensors, 25(13), 4166. https://doi.org/10.3390/s25134166

[12] Huang, S., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. Npj Digital Medicine, 3(1), 136. https://doi.org/10.1038/s41746-020-00341-z

[13] Hu L-S, Wang J, Zhang H-M and Huang H-Y (2025) Automated detection of diabetic retinopathy lesions in ultra-widefield fundus images using an attention-augmented YOLOv8 framework. Front. Cell Dev. Biol. 13:1608580. doi: 10.3389/fcell.2025.1608580

[14] Telçeken, M. (2025). A new hybrid vision transformer model for medical image classifications: Attention mechanism deepened by HOG. Biomedical Signal Processing and Control, 114, 109381. https://doi.org/10.1016/j.bspc.2025.109381

[15] Szanto, D., Erekat, A., Woods, B., Wang, J., Garvin, M., Johnson, B., Kardon, R., Wall, M., Linton, E., & Kupersmith, M. J. (2026). Multimodal deep learning differentiates papilledema and Non-Arteritic anterior ischemic optic neuropathy from healthy eyes. Investigative Ophthalmology & Visual Science, 67(1), 12. https://doi.org/10.1167/iovs.67.1.12

[16] Krones, F., Marikkar, U., Parsons, G., Szmul, A., & Mahdi, A. (2024). Review of multimodal machine learning approaches in healthcare. Information Fusion, 114, 102690. https://doi.org/10.1016/j.inffus.2024.102690

[17] Marouf, A. A., Mottalib, M. M., Ridi, S. S., Jafarullah, O., Rokne, J., & Alhajj, R. (2025). Eye-XAI: an explainable artificial intelligence approach for eye disease detection using symptom analysis. BMC Medical Informatics and Decision Making, 25(1), 433. https://doi.org/10.1186/s12911-025-03253-8

[18] Fang, C., Ma, N., & Qian, L. (2026). Multi-Task Transformer framework and Radiomic signatures for Multi-Lesion segmentation, detection, and grading in diabetic retinopathy. Photodiagnosis and Photodynamic Therapy, 105455. https://doi.org/10.1016/j.pdpdt.2026.105455”

[19] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 53. https://doi.org/10.1186/s40537-021-00444-8”

[20] Ahmed, F., Naz, N. S., Khan, S., Rehman, A. U., Ismael, W. M., & Khan, M. A. (2026). Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. BMC Medical Imaging, 26(1), 37. https://doi.org/10.1186/s12880-025-02118-w