Investigations of Explainable AI based solutions for Diabetic Retinopathy diagnosis

Rysha M Rafeeque, Lohesh M, Rathna Sabapathy P, Esvar Ram Kumar P, and Dhivya S[1][0000−0002−4192−797]

Shiv Nadar University Chennai, Kalavakkam, Tamil Nadu, India
rysha21110088@snuchennai.edu.in, lohesh21110290@snuchennai.edu.in, rathna21110053@snuchennai.edu.in, esvar21110084@snuchennai.edu.in,dhivyas@snuchennai.edu.in

Abstract. This review explores the intersection of explainable artificial intelligence (XAI) and diabetic retinopathy (DR), emphasizing how transparency in AI models can enhance clinical decision-making and patient outcomes. The paper examines various XAI techniques, such as feature attribution, local interpretable model-agnostic explanations (LIME), and SHapley Additive exPlanations (SHAP), and their application in DR diagnosis and prognosis. Through a synthesis of current research, we highlight the advantages of incorporating explainability into AI systems, including improved model trustworthiness and actionable insights for clinicians. Challenges related to interpretability, model complexity, and data privacy are also discussed. By analyzing existing methodologies and their impact on DR management, this review aims to provide a comprehensive understanding of how XAI can drive advancements in diabetic retinopathy care. Ultimately, the paper seeks to guide future research directions and foster the development of more transparent and reliable AI tools in healthcare.

1 Introduction

Diabetic Retinopathy is a prominent cause of permanent blindness for all individuals.It is caused due to diabetes, a long-term illness that interferes with our body's average capacity to digest food. Most of our foods are broken down into glucose and enter our bloodstream. When blood sugar levels rise, our pancreas is pushed to secrete insulin. Insulin is the element that permits blood glucose to enter our body's cells and then be used as food. Whenever a person develops diabetes, the body either does not create enough insulin or does not utilize it that well. There is more blood glucose when insufficient insulin or cells stop producing insulin. Complications of diabetes include diabetic retinopathy (eye damage), neuropathy (nerve damage), nephropathy (kidney disease), cardiomyopathy (heart problems) etc.
So Diabetic Retinopathy treatment must be efficient and hasten. An AI system can reduce doctor workload in terms of time and effort and improve the speed and accuracy. Moreover incorporating this AI can

build solutions that can reach many communities and solve arduous problems. The most naïve approach is gathering the dataset and classifying it through ML/DL techniques. This process of classification is very abstract technically a black box. The predicament over here is the ML/DL model classifies a person as DR/Non DR without any explanation and interpretation. But Explainability is essential for gaining patients' and physicians' trust in the DR diagnosis, why model our model takes this decision, it is right? will it convince physicians?. Figure 1 illustrates the need for explainable AI.Explainable AI(XAI) models attempt to provide users with a more specific explanation for computational outcomes, promoting transparency in AI decision-making.So in this paper we discuss the technical advancements happened in this field over half decade.

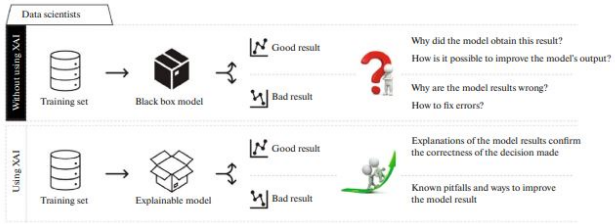

Fig. 1. Schematic representation of the proposed model

2 Datasets

Name of Dataset	**Resolution**	**Sample Size**	**Public**
Kaggle DR Detection	1024x1024 pixels (varies)	88,702 fundus images	Yes
EyePACS	1800x1800 pixels (varies)	88,702 fundus images	Yes
Messidor	2000x2000 pixels	1,200 images	Yes
DRIONS-DB	2000x2000 pixels	110 images	Yes
RetinaPro	2048x1536 pixels	1,000 images	No
STARE	700x605 pixels	400 images	Yes
RiteAid	1920x1080 pixels	1,000 images	No
DDR	1024x1024 pixels	7,000 images	Yes
IDRiD	4288x2848 pixels	81,000 images	Yes
ACDI	2048x2048 pixels	1,000 images	Yes
OCT-DR	512x512 pixels	2,500 images	Yes

2.1 Data Preprocessing

Preprocessing is the preliminary process to curate the data for any missing values or to remove information which does not contribute towards the classification. The feature space is then scaled and normalised to bring it into a range for the further process. While employing the QML, the training and the test data are experimented using a quantum system, which involves only confined qubits. Hence the dimension of the whole data must be reduced that of the original dataset. The underlying principle of Principal Component Analysis (PCA) assists in the feature reduction and also helps in the visualisation of the data. The ability to spot patterns in data and point out their similarities and variations results from this. It seems to be able to condense data with little to no information loss. In this work, to reduce the processing of the features, the dimensions of the features are reduced using PCA and they are visualised in the Figure 2.

2.2 Data split, translation and Execution

The dataset is divided into training and the test set, which are further moved to the quantum based simulator. For simulating the algorithm in the quantum environment, a typical data point must be transformed to a quantum data point. The transformation is achieved with the circuit and further the corresponding labels for the transformed data are obtained using measurements with conventional value of (-1,1). The quantum Support Vector Machine (QSVM) is explored on the WDBC dataset which involves a quantum environment, whereas the other SVM classifier is trained using a classical computer system. The conventional SVM utilise the SVC with different kernels. With respect to the quantum SVM, the quantum kernel is utilized for execution. The quantum states are converted through the quantum feature maps for the kick start of the quantum kernel. The training process is similar to the conventional approach once the kernel matrix is constructed by finding the inner product of the feature maps. The local system or the google collaboratory is connected to the IBM Quantum experience through an IMBid. The developed model is simulated, and its performance is evaluated on the quantum computer, Several parameters are tuned such that the loses are minimized.

3 Literature Survey

Diabetic retinopathy (DR) is a critical complication of diabetes, characterized by progressive damage to the retinal blood vessels, which can lead to vision loss and blindness if left untreated. The diagnosis of DR traditionally relied on manual examination of retinal images by trained ophthalmologists. This process, while effective, is time-consuming and prone to subjectivity, leading to variability in diagnoses, particularly in the early stages of the disease when retinal changes are subtle. To overcome these challenges, Artificial Intelligence (AI) and deep learning

techniques have emerged as powerful tools in the automated detection and grading of DR, offering the potential for more consistent, accurate, and efficient diagnosis.

3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are at the forefront of AI applications in image analysis and are particularly well-suited for medical image processing, including the diagnosis of DR. CNNs are inspired by the human visual system and are composed of multiple layers that automatically learn to extract and hierarchically process features from input images. In the context of DR diagnosis, CNNs are used to analyze retinal fundus images, detecting critical pathological features such as microaneurysms, hemorrhages, exudates, and cotton wool spots.

Studies have demonstrated that CNNs can achieve remarkable diagnostic performance, often surpassing human experts in both speed and accuracy. For instance, CNN-based models have been reported to reach precision rates as high as 0.970 and sensitivity rates of 0.980, making them highly effective in distinguishing between the various stages of DR — ranging from mild non-proliferative DR to severe proliferative DR. These models are particularly valued for their ability to identify early signs of DR, thereby enabling timely interventions that can prevent further progression and vision loss.

However, the primary drawback of CNNs lies in their “black-box" nature. The internal workings of these models are not easily interpretable, which poses a significant challenge in medical applications where the rationale behind a diagnosis is as important as the diagnosis itself. This has led to a growing emphasis on Explainable AI (XAI) techniques that can shed light on the decision-making processes of CNNs, making their outputs more transparent and trustworthy to clinicians.

3.2 Transformers

While CNNs are predominantly used in image analysis, the Transformer architecture, originally developed for sequence modelling tasks in natural language processing, has been increasingly adapted for medical image analysis, including DR diagnosis. Transformers utilize self-attention mechanisms that allow them to capture long-range dependencies and interactions within the data, making them adept at handling complex image data and integrating global contextual information.

In DR diagnosis, transformers have shown promise in enhancing model interpretability and flexibility. Unlike CNNs, which focus on local features through convolution operations, transformers can consider the entire image context, enabling the detection of subtle, distributed patterns that might indicate early stages of DR. This global perspective is particularly valuable when analyzing retinal images, where the spatial relationships between different regions of the retina can provide crucial insights into the disease’s progression.

Emerging studies suggest that transformers can outperform traditional CNNs in certain DR diagnostic tasks by leveraging their ability to integrate both local and global information more effectively. Additionally, transformers' inherent interpretability through attention mechanisms offers a pathway to more transparent AI models, which can be crucial in clinical settings where understanding the decision-making process is essential for patient safety and care.

3.3 Advanced CNN Architectures and Ensemble Methods

Recent advancements in CNN architectures, such as ResNet (Residual Networks), VGG (Visual Geometry Group), and MobileNet, have further enhanced the diagnostic capabilities of AI models in DR. ResNet utilizes skip connections that help mitigate the vanishing gradient problem, allowing the network to train deeper layers and thus capture more complex features. VGG networks, with their use of small convolutional filters, excel at capturing detailed visual features, while MobileNet is optimized for efficiency, making it suitable for deployment in resource-constrained environments where computational power is limited.
Ensemble methods, which combine the predictions of multiple models, have also been applied to DR diagnosis with considerable success. These methods aggregate the strengths of different models, such as CNNs and transformers, to produce more robust and accurate predictions. For example, an ensemble model might integrate the outputs of several CNNs, each trained to focus on different aspects of the retinal image, resulting in a final prediction that is more reliable than any single model alone. This approach not only improves diagnostic accuracy but also enhances the model's generalizability to diverse patient populations and varying imaging conditions.

4 Integration of Explainable AI (XAI) Techniques

The integration of Explainable AI (XAI) techniques into DR diagnosis is critical for ensuring that AI-driven models are not only accurate but also interpretable by clinicians. XAI methods aim to demystify the decision-making processes of complex AI models, providing insights into how and why specific predictions are made. This transparency is especially important in healthcare, where understanding the underlying reasoning behind a diagnosis can significantly impact clinical decision-making and patient outcomes.

SHAP (SHapley Additive exPlanations) SHAP is a powerful XAI technique based on cooperative game theory that assigns a Shapley value to each feature, representing its contribution to the model's output. In the context of DR diagnosis, it can be used to determine which

features of retinal images—such as the presence of specific lesions or abnormal blood vessels—are most influential in the AI model's classification decisions.
SHAP provides both global and local explanations, offering a comprehensive understanding of the model's behavior across the entire dataset as well as insights into individual predictions. This dual capability makes it particularly valuable in clinical settings, where it is crucial to validate AI outputs against established medical knowledge.

Grad-CAM (Gradient-weighted Class Activation Mapping)
Grad-CAM is a visualization-based XAI technique that generates heatmaps over input images, highlighting the regions that contributed most to the model's decision. It does so by computing the gradients of the target class (e.g., a specific DR stage) with respect to the feature maps in the final convolutional layer, producing a coarse localization map that can be overlaid on the original image.
In DR diagnosis, it is particularly useful for providing visual explanations that can be directly compared with human interpretations.

LIME (Local Interpretable Model-agnostic Explanations)
LIME is an XAI technique that creates locally interpretable surrogate models to approximate the behavior of complex models near a specific prediction. By perturbing the input data and observing changes in the model's output, it identifies which features are most important for the model's decision in a particular instance.
In DR diagnosis, it can be used to generate explanations that highlight the contribution of specific image features (such as particular regions of the retina) to the model's prediction. This technique is model-agnostic, meaning it can be applied to any black-box model, including CNNs and transformers.

Saliency Maps Saliency maps are a popular explainability method used to highlight regions in input images that have the most influence on the model's prediction. For DR diagnosis, saliency maps can be used to visualize which areas of the retinal image (e.g., regions with microaneurysms or hemorrhages) the model focused on when making its prediction.

Rule-based Systems In addition to model-agnostic techniques like LIME and SHAP, some researchers have explored incorporating rule-based systems into AI models for DR diagnosis. These systems use predefined rules derived from clinical guidelines to make decisions, providing a more interpretable framework for diagnosis.

4.1 Limitations of XAI Approaches

Despite their strengths, each XAI technique has limitations that must be considered when applying them to DR diagnosis:

- **SHAP:** While SHAP provides valuable insights into feature importance, it is computationally intensive, especially for complex models or large datasets.
- **Grad-CAM:** Grad-CAM is dependent on the model's internal structure, particularly the convolutional layers, making it less effective with other types of models like transformers.
- **LIME:** LIME's reliance on perturbations may not always accurately reflect the true importance of features in the model's decision-making process.

5 Evaluation Metrics

In the context of explainable AI for diabetic retinopathy, the following evaluation metrics are frequently utilized to gauge model performance and interpretability:

- **Accuracy**: Measures the proportion of correctly classified images or cases out of the total number of images or cases. While straightforward, accuracy alone might not suffice in imbalanced datasets where one class significantly outnumbers another.
- **Precision**: Calculates the ratio of true positive predictions to the total predicted positives. In diabetic retinopathy, high precision is crucial to minimize false positives, ensuring that only those cases with high confidence are flagged as positive.
- **Recall (Sensitivity)**: Represents the ratio of true positive predictions to the total actual positives. For diabetic retinopathy, high recall ensures that as many true positive cases as possible are detected, which is critical for early diagnosis and treatment.
- **F1 Score**: The harmonic mean of precision and recall. In the context of diabetic retinopathy, the F1 score balances the trade-off between precision and recall, particularly useful in handling class imbalances where the number of positive cases may be limited.
- **ROC-AUC**: Evaluates the model's ability to distinguish between positive and negative cases across various threshold settings. The ROC-AUC is essential for understanding the trade-offs between true positive and false positive rates, helping in assessing the model's robustness in detecting diabetic retinopathy.
- **Mean Squared Error (MSE)**: Although more common in regression tasks, MSE can be used in evaluating models that output continuous scores or probabilities, such as risk scores associated with diabetic retinopathy.
- **Confusion Matrix**: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives, offering insights into the model's performance across different categories and its error patterns.
- **Explainability Metrics**: Metrics such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are used to evaluate how well the model's predictions can be understood and interpreted. In diabetic retinopathy, these metrics help in elucidating how the AI model arrives at its decisions, which is crucial for clinical trust and actionable insights.

6 Applications

The evaluation metrics are applied in various practical scenarios involving explainable AI in diabetic retinopathy:

- **Predictive Modeling**: AI models are developed to predict the likelihood of diabetic retinopathy progression or occurrence based on retinal images. Metrics like accuracy, ROC-AUC, and F1 score are employed to assess model performance and ensure reliable predictions.
- **Diagnosis Assistance**: XAI models aid ophthalmologists in diagnosing diabetic retinopathy by providing not only predictions but also explanations for those predictions. Precision and recall are used to ensure the model's effectiveness in detecting true cases of the disease.
- **Risk Stratification**: Models predict the risk levels of patients developing severe diabetic retinopathy. Metrics such as MSE and ROC-AUC help in evaluating the accuracy of these risk predictions.
- **Treatment Decision Support**: By explaining model predictions, XAI tools support decision-making for treatment plans. Explainability metrics like LIME and SHAP help clinicians understand which features or aspects of the retinal images are influencing the model's recommendations.
- **Educational Tools**: XAI models can serve as educational resources for training healthcare professionals by providing clear explanations of how different features impact predictions, thereby enhancing their understanding of diabetic retinopathy.
- **Monitoring and Reporting**: Continuous monitoring of model performance using metrics such as confusion matrix and F1 score ensures that the AI system remains accurate and reliable over time, adapting to any changes in data patterns.
- **Compliance and Transparency**: Explainability metrics are crucial for meeting regulatory requirements and ensuring transparency in AI applications in healthcare. They help in building trust with patients and stakeholders by making model decisions understandable and accountable.

7 Challenges and Future Directions in Data Science

7.1 Challenges

1. **Data Quality and Integrity:** Ensuring high-quality data remains a fundamental challenge. Issues such as missing values and data noise can significantly impair model performance. Effective preprocessing techniques are essential but often require tailored approaches based on the data context.
2. **Scalability and Performance:** With the increasing volume and complexity of data, traditional algorithms may struggle with scalability. There is a critical need for more efficient algorithms and computational frameworks, particularly in high-dimensional areas like image analysis.

3. **Model Interpretability:** Complex models, especially deep learning networks, often operate as "black boxes," lacking transparency. Developing methods that provide clear, interpretable insights without compromising model accuracy is vital for gaining stakeholder trust, particularly in sensitive applications.
4. **Bias and Fairness:** Addressing biases in data and ensuring fair outcomes is crucial to avoid perpetuating existing inequalities. This challenge is particularly pressing in fields such as criminal justice and recruitment, where fairness is paramount.
5. **Integration of Multimodal Data:** The integration of diverse data sources, including text, images, and sensor data, necessitates advanced techniques for feature extraction and data alignment.
6. **Data Privacy and Security:** Complying with stringent regulations like GDPR while maintaining data privacy and security poses an ongoing challenge. Ensuring robust data governance practices is essential.
7. **Reproducibility and Standardization:** The lack of standardized methodologies for data preprocessing and model evaluation can hinder research reproducibility. Establishing and adhering to best practices is necessary to enhance the credibility and reliability of scientific findings.

7.2 Future Directions

1. **Advancements in Explainable AI (XAI):** Future research should focus on developing techniques that enhance model interpretability while maintaining high performance. This is particularly crucial in sectors where accountability and transparency are required.
2. **Automated Machine Learning (AutoML):** Enhancing AutoML platforms to optimize for not only performance but also interpretability and fairness will democratize access to machine learning technologies, making them more accessible to non-experts.
3. **Federated Learning and Privacy-Preserving Techniques:** Exploring federated learning and differential privacy offers promising solutions for privacy concerns while enabling decentralized model training and data analysis.
4. **Ethical AI Frameworks:** Developing comprehensive ethical guidelines for AI development and deployment is essential to ensure fairness, accountability, and transparency in AI systems.
5. **Real-Time Analytics and Streaming Data:** Advancing analytical frameworks to handle real-time and streaming data will be critical for deriving actionable insights from continuous data sources.
6. **Interdisciplinary Collaboration:** Strengthening collaborations between data scientists, domain experts, and ethicists will ensure that data-driven solutions are both contextually relevant and socially responsible.
7. **Sustainable Data Practices:** Investigating energy-efficient algorithms and assessing the environmental impact of data collection practices will support the development of sustainable data science methodologies.

8. **Enhanced Data Literacy:** Promoting data literacy across various sectors will empower individuals to effectively understand, interpret, and utilize data insights, fostering more informed decision-making.
9. **Innovative Data Sources:** Leveraging emerging data sources, such as Internet of Things (IoT) devices and social media, has the potential to provide richer insights and enhance model robustness.
10. **Longitudinal Studies and Time-Series Analysis:** Prioritizing longitudinal studies and advanced time-series analysis techniques will improve our understanding of temporal dynamics and trends over time.

8 Conclusion

Diabetic retinopathy (DR) is a critical complication of diabetes that can lead to vision loss if not diagnosed and treated in time. The traditional manual examination of retinal images by ophthalmologists, while effective, is time-consuming and prone to subjectivity. The advent of artificial intelligence (AI) and deep learning has revolutionized DR diagnosis, offering more accurate, efficient, and consistent outcomes.

Convolutional Neural Networks (CNNs) have demonstrated high diagnostic performance in identifying DR stages, often surpassing human experts. Despite their success, CNNs have limitations, particularly their lack of interpretability. To address this, Explainable AI (XAI) techniques such as SHAP, Grad-CAM, LIME, and saliency maps have emerged, enhancing transparency and trust in AI-driven diagnostics. Meanwhile, the introduction of Transformer models and ensemble methods further improves diagnostic accuracy and generalization across different patient populations.

While AI models show remarkable potential in automating DR diagnosis, the integration of XAI techniques is critical to making these systems interpretable and trustworthy for clinical use. By providing insights into model decision-making processes, these explainability techniques help clinicians better understand and trust AI predictions, fostering the adoption of AI tools in healthcare.

Nevertheless, challenges remain in balancing model accuracy with interpretability. Future research must continue to refine AI models and XAI methods to ensure that AI-driven DR diagnosis is both accurate and clinically useful. A collaborative approach involving AI researchers and healthcare professionals will be crucial in achieving this goal, ultimately enabling more effective and scalable DR diagnosis systems.

References

1. Obayya, M., Nemri, N., Nour, M.K., Al Duhayyim, M., Mohsen, H., Rizwanullah, M., Zamani, A.S., & Motwakel, A. (2022). Explainable artificial intelligence enabled TeleOphthalmology for diabetic retinopathy grading and classification. *Applied Sciences*, 12(17), 8749. MDPI.

2. Averkin, A.N., Volkov, E.N., & Yarushev, S.A. (2024). Explainable artificial intelligence in deep learning neural nets-based digital images analysis. *Journal of Computer and Systems Sciences International*, 63(1), 175–203. Springer.
3. Vasireddi, H.K., Devi, K.S., & Reddy, G.N.V.R. (2024). DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment. *Arabian Journal for Science and Engineering*, 1–19. Springer.
4. Quellec, G., Al Hajj, H., Lamard, M., Conze, P.H., Massin, P., & Cochener, B. (2021). ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis*, 72, 102118. Elsevier.
5. Chetoui, M., & Akhloufi, M.A. (2020). Explainable Diabetic Retinopathy using EfficientNET. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1966-1969). IEEE. doi:10.1109/EMBC44109.2020.9175664.
6. Niu, Y., Gu, L., Zhao, Y., & Lu, F. (2021). Explainable diabetic retinopathy detection and retinal image generation. *IEEE Journal of Biomedical and Health Informatics*, 26(1), 44–55. IEEE.
7. Yagin, F.H., Yasar, S., Gormez, Y., Yagin, B., Pinar, A., Alkhateeb, A., & Ardigò, L.P. (2023). Explainable artificial intelligence paves the way in precision diagnostics and biomarker discovery for the subclass of diabetic retinopathy in type 2 diabetics. *Metabolites*, 13(12), 1204. MDPI.
8. Lalithadevi, B., & Krishnaveni, S. (2024). Diabetic retinopathy detection and severity classification using optimized deep learning with explainable AI technique. *Multimedia Tools and Applications*, 1–65. Springer.
9. Alghamdi, H.S. (2022). Towards Explainable Deep Neural Networks for the Automatic Detection of Diabetic Retinopathy. *Applied Sciences*, 12(19), Article 9435. MDPI. doi:10.3390/app12199435.
10. Mridha, K., Wang, M., & Zhang, L. (2024). AI-Driven Diagnostics in Ophthalmology: Tailored Deep Learning Models for Diabetic Retinopathy with XAI Insights. In *Proceedings of the 16th International Conference* (Vol. 101, pp. 73–82).