LAUTECH Journal of Engineering and Technology 19 (1) 2025: 1-14
10.36108/laujet/5202.91.0110

Explainable Ensemble Deep Learning Model for Predicting Diabetic Retinopathy Based on APTOS 2019 Eye Pack Dataset

[1*]Folorunsho O, [2]Akinsanya S. E., [3]Fagbuagun O. A., [4]Mogaji S. A. and [5]Raji S. K.

[1,2,3,4,5]Department of Computer Science, Federal University Oye Ekiti, Nigeria
[1]olaiya.folorunsho@fuoye.edu.ng [2]seye.akinsanya@fuoye.edu.ng, [3]ojo.fagbuagun@fuoye.edu.ng,
[4]Stephen.mogaji@fuoye.edu.ng [5] Sobur.raji.172012@fuoye.edu.ng

Article Info

Article history:

Received: Jan. 11, 2025
Revised: Jan. 22, 2025
Accepted: Jan. 23, 2025

Keywords:

Classification, Deep Learning,
Diabetic Retinopathy,
Interpretability,
Ensemble Model.

Corresponding Author:
olaiya.folorunsho@fuoye.edu.ng
+234(0)8035778999

ABSTRACT

Detection of diabetic retinopathy (DR) as early as possible is vital in mitigating the complicated issues associated with the disease. Recent advances in artificial intelligence (AI), particularly deep learning (DL) techniques, have led to an appreciable increase in the accuracy of predicting various disease classes. However, the challenge of AI models is the difficulty in providing insights into how and why a model arrives in attaining decision-making to facilitate trust and adoption in clinical settings. Therefore, this study aimed to enhance the detection rate of DR and explain the significant regions on the image for the model's overall performance. This study utilised Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Simple Recurrent Neural Networks (SRNN), and XGBoost in an ensemble model (EM). Specifically, Shapley Additive exPlanations (SHAP), a popular Explainable Artificial Intelligence (XAI) technique was utilised to identify and provide insights into which parts of the image's features contribute to the model's overall performance. After a series of experiments using the APTOS 2019 eye pack dataset collected from the Kaggle repository to evaluate the performance of CNN, LSTM, SRNN, and XGBoost. The EM outperformed all the other models with 95.63% accuracy, 97.79% precision, 93.64% recall rate, 98.79% F1-score and 97.75% AUC score. Also, SHAP analysis revealed significant regions on the image that influenced predictions, thus showing how important interpretability was for the model. The results imply that the ensemble DL, particularly with XGBoost, enhances the detection of DR, thereby improving the efficiency of screening tests and supporting personalised treatment plans in clinical practice through integrating these advanced models with XAI tools, creating trust towards automated diagnostic systems.

INTRODUCTION

The eye's retina depends on blood vessels to deliver oxygen and nutrients. The situation gets worse for such patients since raised blood sugars contained in diabetes mellitus can damage the blood vessels (Dubey and Lohiya, 2021). Prolonged high glucose levels weaken the retina's blood vessels, causing fluid leakage, bulging, or the growth of abnormal blood vessels (Donthula and Daigavane, 2024). These changes can result in vision loss and progress into diabetic retinopathy (DR) (Mehboob *et al.*, 2022). Research has shown that DR affects approximately one in every three of the over 463 million people living with diabetes worldwide (IDF, 2023). Men are slightly more likely to develop DR, with an estimated prevalence of 9.0% compared to 7.9% in women (Mohanty *et al.*, 2023). The projection of those affected by DR in the years 2030 and 2045 is estimated to be approximately 643 million and 783 million people, respectively, with

India and China leading (Kumar *et al*., 2024). The most extreme form of DR often shows no symptoms in its early stage. It is, therefore, essential to have retinal screening performed regularly. The methods of diagnosis include licensed practitioners who are generally ophthalmologists meticulously scrutinising the retina under a bright light, which is allowed after anaesthesia of the eye muscle controlling the lens. It is essential to identify the condition in good time to stop further vision impairment in the patients. There are two stages in the development of DR: Non-proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). The NPDR, the early stage type, presents stage one symptoms such as micro-neurism or the development of small bulges in the retina blood vessels, which can lead to fluid accumulation and diffusion vision. If untreated, NPDR can progress to PDR, a more severe stage involving chronic damage and significant vision impairment, highlighting the importance of timely treatment and routine eye examinations (Yadav *et al*., 2021). Figure 1 highlights DR's progression from NPDR to PDR stages.

Computer-Aided Diagnostic Systems (CAD) play an important role in health care systems, as such cost-effective procedures work for various diseases such as DR using colour fundus images (Memari *et al*., 2020). Those systems can rapidly determine such features as the segmentation of the vessel and some characteristics of the optic disc, which assist physicians in making decisions concerning diagnosis and treatment. Medical imaging now heavily relies on advanced tools such as deep learning (DL), a sophisticated subset of machine learning. One of the significant challenges in applying deep learning approaches for DR classification is determining the most effective Convolutional Neural Network (CNN) model for both binary and multi-class classification (Sarki *et al*., 2021). In binary classification, normal retinas are distinguished from diseased ones, whereas multi-class classification involves categorising up to five stages of DR progression (Adriman *et al*., 2021; Dai *et al*., 2024). Research indicates that the use of combined and preprocessed data has a significant impact on improving classification accuracy.

Ensemble Modelling (EM) improves the prediction response by gathering the results obtained from various models. Examples in this category include the Bagging method (e.g., Random Forest), the Boosting method (e.g., AdaBoost), Stacking and voting. These approaches use various models to capture complex datasets for the problem at hand (Macsik *et al*., 2024). Despite significant advancements in medical imaging and deep learning, DR detection remains challenging, especially in translating AI-driven models into real-world healthcare (Yao *et al*., 2024). While ML models show high diagnostic accuracy, their limited interpretability hinders adoption (Antoniadi *et al*., 2021). Medical professionals require diagnostic outputs and insights into the reasoning behind predictions. Without transparency, healthcare providers are reluctant to trust AI models in critical cases like DR (Tucci *et al*., 2022).

Explainable Artificial Intelligence (XAI) enhances transparency in AI systems, which is critical for healthcare applications. Two widely used model-agnostic techniques in XAI are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). These methods provide interpretable explanations for predictions, thereby improving trust and dependability in AI models (Arrieta *et al*., 2020; Man and Chan, 2021). The two XAI techniques rely on a model's internal structure and can be employed to ensure informed decision-making in fields where AI has significant impacts, such as medical diagnosis and finance.

Explainability is crucial in high-stakes medical scenarios, where understanding model predictions is central to decision-making (Kovalchuk *et al*., 2022). Additionally, AI systems often need to provide a comparative diagnostic rationale, leaving clinicians unable to explain clearly to patients (Kaur *et al*., 2020). Variability in disease states and imaging equipment further complicates model development, as deep learning models trained on controlled datasets may perform inconsistently on external data (Zhang *et al*., 2022).

This study proposes an EM model with XAI to predict DR among patients affected with diabetes and subsequently highlight critical image regions influencing predictions. More specifically, the ensemble approach combining CNN, long short-term memory (LSTM), simple recurrent neural networks (SRNN), and XGBoost could enhance DR detection accuracy and reliability (Sikder *et al*., 2021). Integrating XAI into these models is vital for achieving clinically actionable and interpretable outcomes.

RELATED WORKS

Tymchenko *et al*. (2020) introduced a CNN-based approach for DR detection utilising a multistage transfer learning technique. Their method was adapted to the APTOS 2019 Blindness Detection Dataset, comprising approximately 13,000 fundus images, to predict various DR stages. The model achieved impressive sensitivity and specificity of 0.99 each, with a Kappa score of 0.925466, indicating strong agreement between predicted and actual DR stages. A notable strength of this method is its precision in distinguishing DR stages. However, the study highlighted challenges, including the limited availability of affordable labeled datasets and labeling inconsistencies among specialists, which could compromise the training data quality. Despite these challenges, the CNN-based approach demonstrated high accuracy, advancing the applicability of automated DR detection in medical diagnostics.

Khan *et al*. (2021) developed the VGG-NiN model, which combines features from VGG16, Spatial Pyramid Pooling (SPP), and Network in Network (NiN). The hybrid architecture aimed to enhance classification performance and reduce computational demands for DR stage detection. The VGG-NiN model achieved an accuracy of 79.50%, while a comparison with DenseNet121 demonstrated a significantly higher accuracy of 97.30%. Although DenseNet121 outperformed VGG-NiN in accuracy, the latter offered computational advantages critical for real-time medical applications. The study underscored the importance of designing models that balance accuracy and efficiency, contributing to developing automated DR classification systems optimised for resource-constrained environments.

Mushtaq and Siddiqui (2021) proposed a robust DR detection and classification system using DenseNet-169 architecture, validated on the Diabetic Retinopathy Detection 2015 and APTOS 2019 datasets. The model achieved an accuracy of 78%, demonstrating its capability to classify DR stages. While the accuracy fell short of expectations, the study emphasized the system's potential for real-world application due to its feature reuse mechanism, which reduced computational overhead. This work highlighted the importance of creating scalable, efficient systems for early DR detection, a critical factor in timely treatment and vision restoration.

Bora *et al*. (2021) designed a DL model for assessing DR risk and tailoring screening intervals. Using the EyePACS dataset, the model achieved AUCs of 0.79 (internal validation) and 0.70 (external validation). Performance improvements

were observed with the inclusion of additional risk factors, enhancing the model's ability to predict DR progression. However, external validation highlighted dataset bias and the reliance on specific risk factors, which posed challenges for clinical applicability. Despite these limitations, the study contributed to advancing individualized DR risk prediction and optimising screening strategies.

Dai *et al*. (2021) developed the DeepDR system for DR detection and grading, trained on 466,247 fundus images and tested on 200,136 local and 209,322 external images. The model achieved AUC values between 0.943 and 0.972 for DR grading, showcasing its robustness and flexibility. Designed for near real-time detection, DeepDR proved effective for mass screening and clinical use. While the study emphasized its high performance and potential for clinical integration, it also highlighted the need for further validation to ensure widespread applicability in diverse healthcare settings.

Atwany *et al*. (2022) reviewed DL techniques for DR detection, covering supervised and self-supervised learning methods and emerging Vision Transformer architectures. The review highlighted Vision Transformers' ability to model long-range dependencies and improve classification accuracy. Moreover, self-supervised learning is a strategy to reduce the reliance on large annotated datasets. This work offered valuable insights into the evolving landscape of DR detection, emphasizing advanced methods for enhancing accuracy and efficiency.

Ruamviboonsuk *et al*. (2022) developed a deep learning system for DR screening, achieving 94.7% accuracy, 91.4% sensitivity, and 95.4% specificity on Thailand's national diabetes registry. The study demonstrated the system's effectiveness in community settings, comparable to retinal specialists. However, the authors noted that successful deployment requires addressing socioenvironmental factors, including healthcare localisation and staff education. The research highlighted the potential of AI for DR screening in resource-constrained regions while emphasizing the importance of systemic adaptations for successful implementation.

Tejashwini *et al*. (2022) employed a CNN model for DR detection, achieving 81% accuracy. The study introduced explainable AI techniques to enhance transparency and user trust while addressing privacy concerns through robust data protection measures. These efforts improved the model's acceptance and usability in healthcare environments. The study demonstrated the value of combining accuracy, explainability, and privacy in designing AI-driven medical diagnostic systems. Mohanty *et al*. (2023) explored hybrid models combining VGG16, XGBoost, and DenseNet121 for DR detection using the APTOS 2019 dataset. While the hybrid model achieved 79.50% accuracy, DenseNet121 outperformed it with 97.30% accuracy, highlighting the benefits of deeper architectures for extracting complex image features. The study also introduced an app for rapid DR detection, emphasizing the importance of early diagnosis and treatment in reducing vision loss and healthcare costs. Alwakid *et al*. (2023) leveraged image enhancement techniques, including Contrast Limited Adaptive Histogram Equalization (CLAHE) and Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), to improve DR detection using Inception-V3. The model achieved 98.7% accuracy with enhanced images compared to 80.87% without enhancement. This work demonstrated the critical role of image preprocessing in improving model performance and emphasized its potential for accurate and reliable DR detection.

Hussain *et al*. (2023) developed a sophisticated deep-learning ensemble model of ResNet50 and InceptionV3 that automatically detects DR from

retinal fundus images. It was validated against the DIARETDB1 database and obtained a commendable accuracy level of 98.37%, which demonstrates effectiveness in overcoming obstacles of speed and precision associated with DR detection. Dai *et al*. (2024) introduced DeepDR Plus, a system for predicting DR progression and recommending personalized screening intervals. The model showed excellent predictive capability with concordance indices ranging from 0.754 to 0.846. By addressing individual progression risks, DeepDR Plus optimised resource allocation and care efficiency, advancing the integration of personalised medicine in DR management.

The reviewed studies demonstrate significant progress in DR detection and classification using deep learning methods. Key advancements include the integration of novel architectures, transfer learning, image enhancement techniques, and personalised screening strategies. However, challenges such as dataset bias, labeling inconsistencies, privacy and model transparency concerns remain, underscoring the need for continued research to refine these systems for clinical and real-world applications.

METHODOLOGY

This section describes the methodology of the proposed system for detecting DR by incorporating ensemble deep learning models with XAI to make them more explainable.

Data Collection

The Asian Pacific Tele-Ophthalmology Society (APTOS) 2019 Eyes Pack dataset used for this study was obtained from Kaggle and comprised of 3,662 fundus images: 1,805 No DR, 370 Mild DR, 999 Moderate DR, 193 Severe DR, and 295 Proliferative DR. Photographic images of both eyes were captured in 224 × 224-pixel resolution.

Pre-processing of Image Data

In the preprocessing stage, the initial multiclass labels in the dataset were transformed into binary labels. The images were then dimensioned to 128×128 pixels, with their pixel values centered within a range from 0 to 1. Further, the data was randomly divided into training and test data sets in the ratio 70:30, respectively, while stratified sampling ensured even distribution across classes within both sets. The Synthetic Minority Oversampling Technique was applied to the issue of class imbalance since the method developed additional samples for the under-represented classes. Further, the Image Data Generator was used to increase the training data and random augmentations, such as rotation, shifting, shearing, zooming, and horizontal flipping. For DR detection with XAI, these preprocessing steps got everything to a properly grounded level for the training and testing of the ensemble DL models.

Classification Models

Classification models are a category of ML algorithms specifically crafted to allocate predefined labels or categories to input data, relying on its distinctive features. These models find widespread application in various tasks, including but not limited to spam detection, image recognition, and sentiment analysis. Below are the classification models applied in this research:

Convolutional Neural Networks

The CNNs are types of neural network architectures designed to execute tasks that handle data in a grid-like fashion, such as images. They have become instrumental in image recognition, object detection, and other computer vision applications. Major constituents of CNNs include convolutional layers, pooling layers, and fully connected layers. A summary of the mathematical foundation that supports CNNs is given as follows:

The convolutional layer serves as the foundational element in a CNN, executing convolution operations on input data using adaptable filters or kernels. Given an input tensor (X) and a filter (W), the convolution operation can be articulated as:

$$(X \times W)_{i,j} = \sum m \sum n \, X_{(i+n)} / W_{m,n} \quad (1)$$

where (i) and (j) denote the output feature map's spatial indices, while (m) and (n) iterate over m and n are over the filter dimensions. Note that the operation between the filter and the local region of the input consists of element-wise multiplication followed by summation.

Long Short-Term Memory

The LSTM is a Recurrent Neural Networks (RNNs) category designed to capture long-term dependencies and address the vanishing gradient problem. Long Short-Term Memory features a more complex architecture than traditional RNNs, incorporating specialized memory cells to store and manage information over time. The forget gate (f_t) dictates which information from the previous cell state (c_{t-1}) should be omitted.

$$f_t = \sigma\,(W_f . [h_t - 1, x_t] + b_f) \quad (2)$$

The input gate (i_t) determines which new information from the input (x_t) should be incorporated into the cell state.

$$i_t = \sigma\,(W_i . [h_t - 1, x_t] + b_i) \quad (3)$$

The output gate (ot) determines which information from the current cell state (c_t) should be utilised to compute the hidden state.

$$o_t = \sigma\,(W_o . [h_t - 1, x_t] + b_o) \quad (4)$$

Where:

(W_f, W_i, W_o) denote weight matrices.

(b_f, b_i, b_o) represent bias vectors.

(σ) corresponds to the sigmoid activation function.

($[h_t - 1, x_t]$) signifies the concatenation of the previous hidden state and the current input.

Recurrent Neural Network

The RNN represents a neural network architecture specifically crafted for processing sequential data, incorporating information from previous time steps. The mathematical expressions for a basic RNN are outlined below:

Given (x_t) as the input at time (t) and (h_t) as the hidden state at time (t), with (W_{ih}) denoting the weight matrix for input-to-hidden connections, (W_{hh}) representing the weight matrix for hidden-to-hidden connections, and (b_h) serving as the bias vector for the hidden layer, the equations governing a simple RNN are:

$$h_t = \tanh(W_{ih} . x_t + W_{hh} . h_{t-1} + b_h) \quad (5)$$

Where:

W_{ih} corresponds to the weight matrix for input-to-hidden connections.

W_{hh} represents the weight matrix for hidden-to-hidden connections.

b_h is the bias vector for the hidden layer.

tanh designates the hyperbolic tangent activation function.

Extreme Gradient Boosting

Extreme Gradient Boosting stands out as a widely used ML algorithm applicable to regression and classification tasks. The mathematical representation of XGBoost involves an objective function, fine-tuned during the training process. For the binary classification task, akin to logistic regression, the objective function is articulated as follows:

In the context of a training dataset(x_i, y_i), where x_i denotes the features and , y_i is the binary label (0 or 1), the XGBoost objective function comprises the

summation of the logistic loss and a regularisation term:

$$\text{Objective} = \sum i + \sum k \quad (6)$$

Where:

$\sum i$ = Summation of the logistic loss

$\sum k$= Summation of the regularisation term

SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) is a framework designed to explain the outcomes of ML models. A method emanating from cooperative game theory equitably distributes contributions to the various features toward making a prediction; SHAP values answer the question: "For a given prediction, what is the contribution of each feature to that prediction?" In ML, SHAP values extend the above concept to feature attributions. They provide a way of unbiasedly distributing the value of the prediction across the input features. Given a prediction, the SHAP values enumerate the average contribution of each feature across all possible coalitions of features. This considers all possible subsets of features in computing the average contribution of each feature to every possible coalition. Consequently, the SHAP values for each feature are computed as the expected value of its marginal contribution across all combinations of the features. The Shapley value computes the contribution of a feature to be the average one.

In a very intuitive and transparent way, SHAP values convey the contribution of every feature toward the model's prediction on a given instance. Large positive SHAP values mean a value has a substantial positive contribution toward the prediction, whereas large negative values mean a strong negative contribution.

The Mathematical model for each of the features i is depicted in equation 3.12.

$$\emptyset_i = \sum_{S \subseteq N\{I\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (7)$$

Where;

N = Set of all features

i = specific feature from the set of all possible features N

S = Subset of N which does not include feature i

|S| = The number of features in subset S

f(S) = The prediction from the model using the subset feature without considering the feature i

$f(S \cup \{i\})$ = The prediction from the model using the subset feature plus the feature i

$\emptyset_i$ = The Shapley value for the feature i

Experimental Setup

The Adam Optimiser was used to train all the models with a learning rate of 0.0001, a batch size of 25, and a maximum of 20 epochs. *ReLU* activation was turned on for hidden layers of the CNN, and finally, the Sigmoid function was activated at the output of the network, which had two hidden layers comprising 32 and 64 neurons, respectively, together with the dense layer. The Sigmoid function was utilised, and the configuration consisted of 64 hidden neurons for the output layer of the LSTM model. The output layer of the SRNN model also employed the Sigmoid function and contained 64 hidden neurons. The damper effects of the various models were measured by keeping the same hyperparameters across the models, which enabled a streamlined evaluation of their effectiveness in the task.

Performance Evaluation

The models were evaluated based on the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) values generated from the confusion matrix. The performance metrics, including accuracy, precision, recall, f1-Score and Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), were calculated from the values derived from the confusion matrix.

1. *Accuracy*

Accuracy determines how effective the model is by checking the amount of the instances which was accurately predicted against the total number of classifications.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (8)$$

2. *Precision*

Precision checks the positive instances amongst the predicted positive instances in the model. The true positive is divided from the true positive and false negative. Whenever the precision score is high then the false positive rate reduces.

$$Precision = \frac{TP}{TP+FN} \quad (9)$$

3. *Recall*

Recall is also referred to as specificity or accepted positive rate. It evaluates how well the model can check the performance of the genuine instances amongst all the genuine instances in the dataset. Whenever the model has more of positive instances then there is an assurance that there is a minute rate of negative instances.

$$Recall = \frac{TP}{TP+FP} \quad (10)$$

4. *F1-Scor***e**

To determine the f1 score of a model, the precision multiplied by the recall which is also multiplied by 2 is divided by adding the precision and recall scores. In mathematical terms, it is expressed as:

$$F_1 = \frac{2\times precision\times recall}{precision+recall} \quad (11)$$

5. *Area Under the Curve-Receiver Operating Characteristic Curve (AUC-ROC)*

AUC-ROC measures the area under the ROC curve which depicts a tradeoff between sensitivity (or recall) and specificity at different classification thresholds. The ROC curve shows how many true positive predictions are made across various threshold points versus how many false alarms were raised.

$$\text{AUC} = \int_0^1 TPR(t)d(FPR(t)) \quad (12)$$

RESULTS AND DISCUSSION

This section presents the results and discusses the DL techniques for detecting DR, the feature contribution analysis for the EM, and the performance comparison of the proposed model with existing models in the literature.

The Classification of the Ensemble Model

The confusion matrices for CNN, LSTM, SRNN, and XGBoost, as shown in Figures 2a–d, were used to derive the performance metrics summarised in Table 1.

Table 1: Comparative Evaluation of Performance of the Ensemble Model

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
CNN	94.17	96.31	92.88	94.56	96.85
LSTM	92.81	95.73	90.69	93.14	96.43
SRNN	92.53	92.80	91.67	92.23	96.25
XGBOOST	95.63	97.79	93.64	98.79	97.75

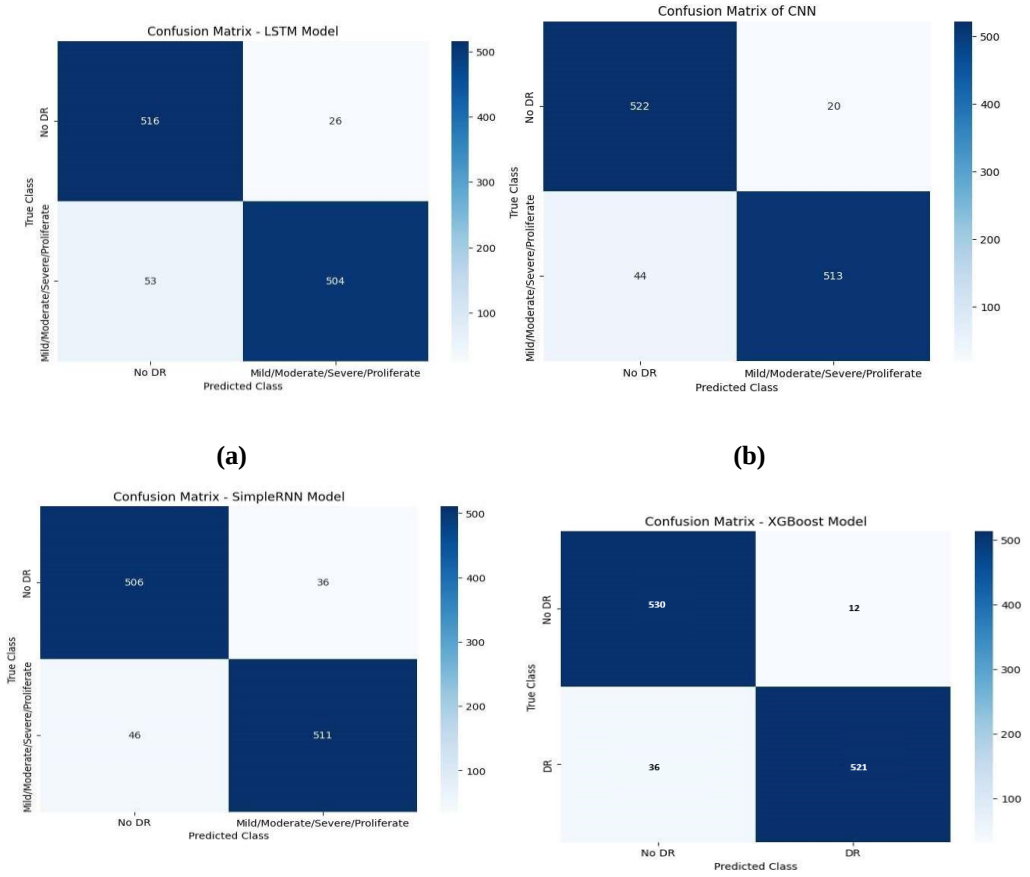

Figure 2: Confusion Matrix of (a) CNN (b) LSTM (c) SRNN (d) XGBoost

The results in Table 1 highlight the performance of the models on the APTOS 2019 Eye Pack Dataset and their effectiveness in classifying DR severity. Among the models, XGBoost achieves the best results across all metrics, with the highest accuracy (95.63%), precision (97.79%), recall (93.64%), F1-score (98.79%), and AUC (97.75%). These metrics demonstrate XGBoost's exceptional ability to classify cases accurately while minimising false positives and false negatives, making it the most reliable and robust model for this task.

The CNN also performs well, achieving an accuracy of 94.17%, precision of 96.31%, recall of 92.88%, F1-score of 94.56%, and AUC of 96.85%. Its high precision and recall reflect its strength in capturing spatial features, such as lesions and abnormalities in retinal images, which are critical for identifying diabetic retinopathy. Although slightly outperformed by XGBoost, CNN remains a strong candidate for image-based classification tasks.

The LSTM model achieves an accuracy of 92.81%, precision of 95.73%, recall of 90.69%, F1-score of 93.14%, and AUC of 96.43%. While its performance is solid, its lower recall compared to CNN and XGBoost indicates that it may miss more positive cases. This could be attributed to LSTM being better suited for sequential data rather than spatial image features, which are critical in this dataset. Finally, the SRNN has the lowest performance among the models, with an accuracy of 92.53%, precision of 92.80%, recall of 91.67%, F1-score of 92.23%, and AUC of 96.25%. Although it performs well, its slightly lower metrics suggest that it is less effective in distinguishing the severity of DR than the other models.

While all models perform strongly, XGBoost is the most effective, followed closely by CNN. Both models are excellent for automating DR detection and classification using the APTOS 2019 dataset.

SHAP Analysis of the Model

After training and testing the model, SHAP analysis was conducted to identify the image regions most influential in predicting outcomes, as shown in Figure 3. Additionally, SHAP analysis highlighted the individual model features that contribute significantly to the predictive power of the EM, as depicted in Figure 6. This analysis reveals the areas within images that strongly impact the model's decisions and clarifies which specific features from each model enhance the ensemble's performance, enhancing interpretability and aiding in refining the model for more accurate DR detection.

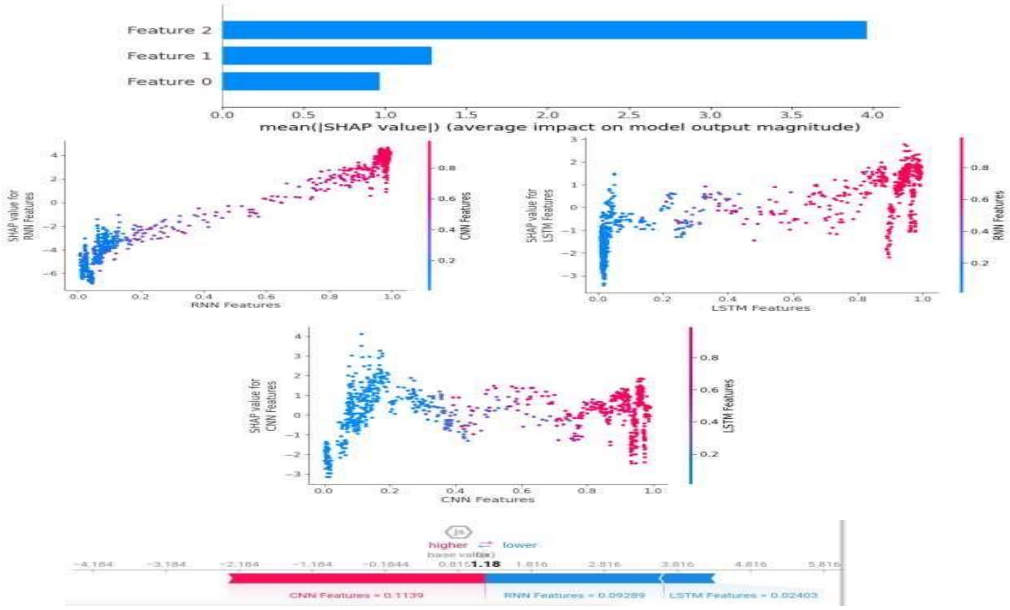

Figure 3: Application of SHAP on the Model to Show the Top Regions Affecting the Model Prediction

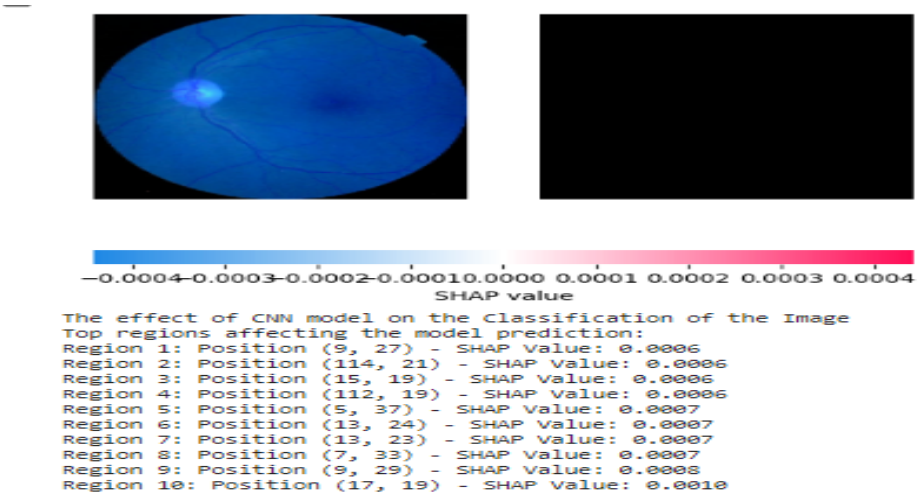

Figure 4: Application of SHAP on the XGBoost Ensemble Model to Interpret the Effect of the Individual

The SHAP analysis results in Figure 4 illustrate the contribution of each feature in the ensemble model to the final predictions. The results indicate that *Feature 2*, extracted from the CNN model, has the most significant impact on the XGBoost ensemble predictions, underscoring the importance of

capturing spatial features through CNNs to enhance performance accuracy. *Feature 1*, derived from the LSTM model, ranks next in importance, though its impact still needs to reach the level of the CNN. Finally, *Feature 0*, contributed by the SRNN model, has the least influence on the ensemble predictions. Although the SRNN model provides informative outputs, its overall contribution is minimal compared to the other models. SHAP analysis thus verifies the value of each model iteration within the ensemble, demonstrating that combining models enhances accuracy, prediction consistency, and reliability.

Performance Comparison of the Proposed Techniques with the Existing Techniques

The comparative analysis between the proposed model and three existing systems is presented in Table 2.

Table 2: Performance Comparison of the Proposed Techniques with the Existing Techniques

Author	Ensemble Model Employed	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Qummar *et al.*, (2019)	Inceptionv3, Xception, Dense121, Dense169	80.80	-	51.50	-	-
Sikder *et al.*, (2021)	VGG16, VGG19, EfficientNetB5, EfficientNetB7, and EfficientNetV2S	94.20	-	-	93.51	-
Bodapati and Balaji (2024)	CNNs, Bidirectional Gated Recurrent Units (Bi-GRUs), Deep Memory Networks (DMNs)	86.22	-	-	-	96.47
Proposed Model	**CNN, SRNN, LSTM**	**95.63**	**97.79**	**93.64**	**98.79**	**97.75**

The comparison was based on the methods employed by the authors and their reported performance metrics, with all authors using the same dataset, APTOS Eye Packs and Stacking Ensemble model. The results, presented in Table 2, show that the proposed model outperformed the systems developed by Qummar *et al*. (2019), Sikder *et al*. (2021), and Bodapati and Balaji (2024). These results demonstrate the effectiveness of the proposed approach in achieving the desired outcomes, particularly by leveraging the classification strengths of XGBoost in combination with architectures such as CNN, LSTM, and SRNN. The EM integrates the strengths of each component, significantly reducing the overall classification error and effectively handling diverse data patterns to enhance DR prediction. By combining multiple learners, the model produces more accurate and generalisable outputs, highlighting the advantage of ensemble techniques over individual classifiers. These findings reinforce the potential of deep learning-oriented ensemble frameworks for medical image classification, underscoring the need for

multi-model approaches to improve accuracy in this domain.

CONCLUSION

This study highlights the substantial progress in DR detection achieved through the use of diverse DL models. The analysis of CNN, LSTM networks, SRNN, and XGBoost has confirmed their effectiveness in diagnosing and classifying DR. XGBoost, in particular, stood out as the most effective model, emphasizing the advantages of ensemble learning techniques in medical image analysis. Additionally, SHAP analysis was instrumental in understanding the models' decision-making processes. By examining how different regions of the images influence predictions, SHAP revealed which features are most critical for diagnosing DR. This interpretability is essential for building trust in ML models, especially in healthcare, where clear explanations of model decisions can aid in clinical decision-making and enhance patient care. The adoption of these sophisticated techniques not only improves diagnostic accuracy but also supports more personalised and efficient screening practices. Understanding and interpreting key features associated with DR progression enables targeted interventions and more effective management strategies. The findings from this research advocate for ongoing innovation and the application of ensemble learning and XAI technologies to advance automated diagnostic systems in healthcare.

REFERENCES

Adriman, R., Muchtar, K., and Maulina, N. (2021). Performance evaluation of binary classification of diabetic retinopathy through deep learning techniques using texture feature. Procedia Computer Science, 179, 88-94.

Alwakid, G., Gouda, W., and Humayun, M. (2023, March). Deep Learning-based prediction of Diabetic Retinopathy using CLAHE and ESRGAN for Enhancement. In Healthcare 11, 863-879.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., and Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences, 11(11), 1-23.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Atwany, M. Z., Sahyoun, A. H., and Yaqub, M. (2022). Deep learning techniques for diabetic retinopathy classification: A survey. IEEE Access, 10, 28642-28655.

Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., ... and Bavishi, P. (2021). Predicting the risk of developing diabetic retinopathy using deep learning. The Lancet Digital Health, 3(1), e10-e19.

Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., ... and Jia, W. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature communications, 12(1), 32-42.

Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., and Jia, W. (2024). A deep learning system for predicting time to progression of diabetic retinopathy. Nature Medicine, 30(2), 584-594.

Donthula, G., and Daigavane, S. (2024). Diabetes Mellitus and Neurovascular Pathology: A Comprehensive Review of Retinal and Brain Lesions. Cureus, 16(10), e70611.

Dubey, A., and Lohiya, S. (2021). Changes in Eyes in a Diabetic Patient. Journal of Pharmaceutical Research International, 33(61A), 480-485.

Hussain, M. M., Shanmugam, P., Moorthi, K., Sakthivelu, U., Rajasekar, A., and Kumar, R. N. (2023). An Ensemble Deep Learning Model for Diabetic Retinopathy Identification. In *2023 9th International Conference on Smart Structures and Systems (ICSSS)*, 1-7.

IDF Diabetes Atlas. Available online: https://diabetesatlas.org/atlas/ninth-edition (accessed on 27 October 2024).

Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., ... and Islam, S. R. (2020). Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives. IEEE Access, 8, 228049-228069.

Kaushik, V., Gessa, L., Kumar, N., and Fernandes, H. (2023). Towards a new biomarker for diabetic retinopathy: exploring RBP3 structure and retinoids binding for functional imaging of eyes in vivo. International Journal of Molecular Sciences, 24(5), 4408-4425.

Khan, Z., Khan, F. G., Khan, A., Rehman, Z. U., Shah, S., Qummar, S., ... and Pack, S. (2021). Diabetic retinopathy detection using VGG-NIN a deep learning architecture. IEEE Access, 9, 61408-61416.

Kovalchuk, S. V., Kopanitsa, G. D., Derevitskii, I. V., Matveev, G. A., and Savitskaya, D. A. (2022). Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. Journal of Biomedical Informatics, 127, 104013-104028.

Macsik, P., Pavlovicova, J., Kajan, S., Goga, J., and Kurilova, V. (2024). Image preprocessing-based ensemble deep learning classification of diabetic retinopathy. IET Image Processing, 18(3), 807-828.

Man, X., and Chan, E. P. (2021). The best way to select features? comparing mda, lime, and shap. The Journal of Financial Data Science, 3(1), 127-139.

Mehboob, A., Akram, M. U., Alghamdi, N. S., and Abdul Salam, A. (2022). A Deep Learning Based Approach for Grading of Diabetic Retinopathy Using Large Fundus Image Dataset. Diagnostics, 12(12), 3084-3098.

Memari, N., Abdollahi, S., Ganzagh, M. M., and Moghbel, M. (2020, September). Computer-assisted diagnosis (CAD) system for Diabetic Retinopathy screening using color fundus images using Deep learning. In 2020 IEEE Student Conference on Research and Development (SCOReD) (pp. 69-73). IEEE.

Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V. C., Karamitsos, I., and Kanavos, A. (2023). Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. Sensors, 23(12), 5726-5744.

Mushtaq, G., and Siddiqui, F. (2021, February). Detection of diabetic retinopathy using deep learning methodology. In IOP conference series: materials science and engineering (Vol. 1070, No. 1, p. 012049). IOP Publishing.

Ruamviboonsuk, P., Tiwari, R., Sayres, R., Nganthavee, V., Hemarat, K., Kongprayoon, A., and Webster, D. R. (2022). Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. The Lancet Digital Health, 4(4), e235-e244.

Sarki, R., Ahmed, K., Wang, H., Zhang, Y., and Wang, K. (2021). Convolutional neural network for multi-class classification of diabetic eye disease. EAI Endorsed Transactions on Scalable Information Systems, 9(4).

Sikder, N., Masud, M., Bairagi, A. K., Arif, A. S. M., Nahid, A. A., and Alhumyani, H. A. (2021).

Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry, 13(4), 670-685.

Tucci, V., Saary, J., and Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review. Journal of Medical Artificial Intelligence, 5, 1-13.

Tejashwini, D., Gaonkar, M. S., Lakshmi, H. D., Mary, A. R., and Madhuri, J. M. (2022). An explainable AI model for diabetic retinopathy detection. International Journal of Innovative Research in Advanced Engineering, (Vol. 9, Issue 8, pp. 306–311).

Tymchenko, B., Marchenko, P., and Spodarets, D. (2020). Deep learning approach to diabetic retinopathy detection. arXiv preprint arXiv:2003.02261.

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083.

Yadav, P., Singh, S. V., Nada, M., and Dahiya, M. (2021). Impact of severity of diabetic retinopathy on quality of life in type 2 Indian diabetic patients. Int J Community Med Public Health, 8, 207-211.

Yao, J., Lim, J., Lim, G. Y. S., Ong, J. C. L., Ke, Y., Tan, T. F., ... and Ting, D. S. W. (2024). Novel artificial intelligence algorithms for diabetic retinopathy and diabetic macular edema. Eye and Vision, 11(1), 1-12.

Zhang, A., Xing, L., Zou, J., and Wu, J. C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. Nature Biomedical Engineering, 6(12), 1330-1345