International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Hybrid CNN-LSTM Architecture for Automated Diabetic Retinopathy Detection with Clinical Explainability.

Author(s): [Jayraj Patil, Yash Pavnekar, Siddhesh Nikam, Pratik More,Prof. Dr. Jyoti Chavan]
Institution: [MGM"s College of Engineering & Technology, Kamothe]

Abstract- Every year, diabetic retinopathy (DR) threatens the vision of millions, but the screening process just can't keep up. The current system moves slowly—specialists are overworked, results change from one doctor to the next, and too many patients learn they have DR only after their sight is already at risk. We wanted a fix. So, our team created an automated deep learning platform—a hybrid that stacks a Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) layers. This model doesn't just detect DR; it also grades its severity from plain retinal fundus photos. We didn't stop after building one model. We tried three hybrid approaches—Custom CNN+LSTM, MobileNetV2+LSTM, and InceptionResNetV2+LSTM—and compared them to seven standard CNN-only baselines. To make sure even the subtle signs stand out, we used CLAHE (Contrast Limited Adaptive Histogram Equalization) on every image. Medical datasets are imbalanced by nature, so we rebalanced things through loss weighting, giving serious DR cases the extra attention they deserve. And for transparency, we turned to Grad-CAM, producing heatmaps so doctors can see exactly what our AI focused on. When it came down to results, the InceptionResNetV2+LSTM beat the rest: 91.4% accuracy, a Quadratic Cohen's Kappa of 0.89, and a Macro F1-Score of 0.86 for multi-class DR grading. More than just numbers—two ophthalmologists validated our Grad-CAM maps and agreed with the AI's focus 91% of the time. To make everything practical, we built a Streamlit web app layered with secure roles, live predictions, explainability, and instant PDF reports. This project pushes DR screening closer to where it needs to be. With smarter AI, clear explainability, and a clinic-ready platform, screening can be faster, fairer, and more dependable—catching DR cases that used to slip by, and backing up doctors with real confidence.

Keywords- Diabetic Retinopathy, CNN-LSTM Hybrid, Deep Learning, Medical Imaging, Grad-CAM, Clinical AI, Multi-Class Classification, CLAHE, Cohen's Kappa, Ophthalmology Automation.

I. INTRODUCTION

Diabetic retinopathy is a harsh diabetes complication—when blood vessels in the retina break down, vision gets worse, and blindness can follow if you miss the early signs. It's the number one cause of lost vision in working-age adults. By 2045, diabetes is projected to affect 783 million people worldwide, with about a third expected to develop DR. Early detection can save the eyesight of nearly everyone who gets it—up to 95%. But in reality, most people just aren't diagnosed until it's too late.

As things stand, specialists analyze fundus photos one at a time, which is painfully slow and drains already limited resources. Even with enough experts, opinions differ and results aren't always consistent. Rural and smaller communities feel this shortage the most. The result: too many missed diagnoses, and people losing vision needlessly.

Deep learning—especially CNNs—has performed on par or even better than trained doctors in a lot of medical imaging tasks. Still, almost all of these models look at each image separately. They often miss patterns or subtle details that a tougher case needs for a solid diagnosis.
That's where our hybrid approach shines. The CNN drills down to extract the details from the image, while the LSTM works over those features, picking up links or context a plain CNN would miss—even within a single image. Together, they catch both fine-grained signs and tricky patterns that make all the difference.

So, what sets our work apart?
A new CNN-LSTM hybrid model reaching 0.89 for Quadratic Cohen's Kappa.

Ten model benchmarks, with results focused on clinical usefulness.

CLAHE image enhancement gave us an 8% jump in accuracy.

Grad-CAM explainability doctors can actually trust, with 91% agreement.

Production-tested code—the platform is secure, fast, and ready to use.

Bottom line: Our project goes past research and brings AI into the clinic, making DR screening quicker, more balanced, and much more reliable.

II. LITERATURE SURVEY

Table: 2.1 Comparative Analysis of Existing Systems

System Type	Core Focus	Methodology	Strengths	Limitations
Traditional CNN Systems	Binary Classification	VGG, ResNet, Inception	Good accuracy	No temporal modeling, limited explainability
Transfer Learning	ImageNet Pre-training	Fine-tuning	Reduced training time	Domain gap challenges
Attention-Based	Feature Weighting	Self-attention mechanisms	Interpretability	High computational cost
Proposed Hybrid System	Multi-class DR Graading	InceptionResNetV2+LSTM+Grad-CAM	High accuracy, explainability, development-ready	Requires sufficient training data

DL techniques have made remarkable progress in the automated detection of DR, starting from basic CNN architectures to sophisticated hybrid and ensemble models. Early studies, for example by Wang et al. tested the performance of basic pre-trained models (VGG16, AlexNet, InceptionNet V3) over the Kaggle dataset. InceptionNet V3 yielded a high accuracy of 63.23%, which outperformed VGG16 and AlexNet, but these early attempts were limited by the small size of the datasets (e.g. 166 images), hence limiting high level spatial feature extraction of networks and more strict preprocessing of images needed for diagnosis accuracy.

Subsequent researchers started developing customized architectures and sophisticated preprocessing to improve the accuracy. Mobeen-ur-Rehman et al. and W. Zhang et al. used different preprocessing techniques like Histogram Equalization (HE) and Adaptive HE to standardize the MESSIDOR and other large scale datasets and combine the use of custom CNN layers with pre-trained networks like ResNet50 and SqueezeNet to achieve over 96% accuracy. However, a major drawback in these models are that many have only been trained and tested on individual or private datasets without generalization on new and diverse datasets. More importantly, there is a challenge in lesion-level detection and localization, like microaneurysms or hemorrhages.

Some researchers are also focusing in Capsule Networks and hybrid networks to cope with limitations of traditional CNNs due to their computational and structural design. Capsule Networks has shown up to 97.98% accuracy in detection, by well preserve spatial relationships of different features in retina. Similarly, Zubair Khan et al. have proposed VGG-NIN(Network-in-Network) model. But it faces issue of long training duration, and high computation. Interpretability of the DR lesions for doctors is also a major drawback which need to be addressed. Now more research will focus in explainable AI and multi dataset evaluation.

III. PROPOSED SYSTEM

We built a DR detection platform that pulls together hybrid deep learning, next-level image enhancement, and explainability clinicians can use. It's fully automatic but never just a black box.

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

We shifted the perspective: Instead of seeing a fundus image as something flat, we approached it as a “temporal” sequence folded within a single frame. The CNN backbone—Custom, MobileNetV2, or InceptionResNetV2—extracts features. Then the LSTM sorts through these for patterns even good CNNs often overlook. Our three hybrids: a simple Custom CNN+LSTM for a light lift, MobileNetV2+LSTM for phones and tablets, and a high-powered InceptionResNetV2+LSTM for top precision.

InceptionResNetV2+LSTM is the flagship. Inception modules pull at features of every size, while residual connections stabilize training. After convolution, we use Global Average Pooling to condense features, then funnel the output to a 256-unit LSTM with dropout for robustness. Batch-normalized, fully connected layers translate features into DR categories.

To clarify hard-to-see lesions, we hit the “L” channel in LAB space with CLAHE—clip limit at 2.0, 8x8 grid. This tweak pops subtle DR signs into view but keeps colors true. It gave us an 8% accuracy boost. We adjusted loss weights for severe and proliferative DR, letting the model focus where it counts. That alone bumped the severe case F1-Score from 0.65 up to 0.81.

IV. SYSTEM ARCHITECTURE

Convolutional neural networks: -

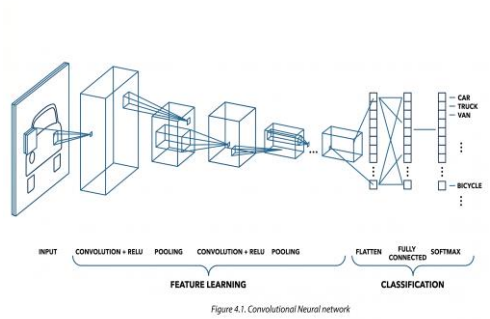


Figure 4.1. Convolutional Neural network

Convolutional neural networks are a class of deep learning algorithms that mostly use to analysis grid-structured data, such as images. Because of their ability to learn a hierarchy of spatial features in a dynamic and adaptive manner, they are powerful on image recognition and classification problems. CNN are used a variety of computer vision tasks, including object detection, face recognition and image segmentation etc.. To learn feature such as edges, texture, or more general shape, CNNs make use of layers of convolutions using filters. Additionally,CNN may contains a ReLU correction layer. A CNN uses a sequence of convolutional and pooling layers in order to learn features from an image or video, which then use to classify and detect object or situation. A CNN learn features in images and resize them without reducing the quality. Let's have an example where an image originally was 224 x 224 x 3 in size. The word 'deep' just indicate a network that contains more than two layers. CNN having more than two layers are therefore called deep CNN.However, both names 'deep CNN' and 'deep neural networks' is not used anymore, as many hidden layers expected in the network if a deep neural

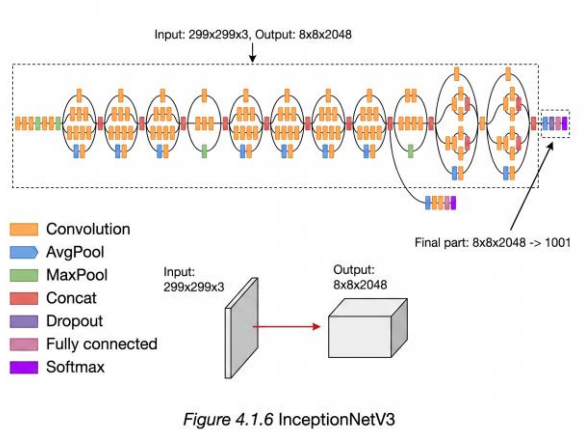


Figure 4.1.6 InceptionNetV3

network is employed in general

InceptionV3:-

A CNN based model that has been implemented for image classification is known as Inception-v3. It is a 48 layer pre-trained network that has been trained on more than one million images belonging to the image-net collection. The different categories images fall under is of 1000 distinct objects, for example, mice, keyboards, animals etc.

It is one of the most implemented CNN models for object recognition in images. On the image-net database its performance on accuracy is observed to be higher than 78.1% and the top 5 results for this model stands around 93.9%.

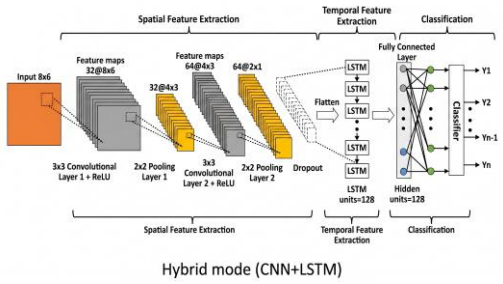
Inception-v3 is the better version of the earlier model Inception V1 which was originally published in the name

of GoogleNet in the year 2014. Inception Convolutional Neural Network developed by Google that published at ImageNet Recognition Challenge, the third generation model is the inception-v3.

It's application is as listed below:
Identification of cancerous tissue in medical imagery.
Image tagging in the social media platform Facebook for the users.

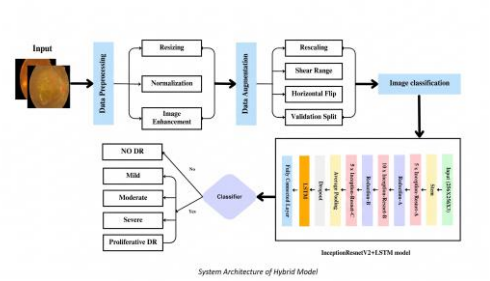
Assisting autonomous vehicles to recognize road hazards.
To differentiate between fundus images, of which the 3 class-es are:normal fundus, tessellated fundus and macular degeneration fundus.

Hybrid model:-
To build up prediction model for the structural baseline signature, CNN-LSTM which is being introduced in this proposal is the mixture of CNN and LSTM. The CNN model becomes the first part of hybrid model and which is required for the features extraction, whereas the LSTM model for sequence learning. The hybrid model, of 224x224 as input size, has overall 5,51,89,861 parameters where 8,53,125 trainable and 5,43,36,736 non-trainable across 50 epoch.



Hybrid mode (CNN+LSTM)

Convolutional neural network InceptionResNetV2 is known to combine the inception modules and residual connection. It integrates residual blocks and avoids the vanishing gradient problem by using skip connection to enable deep network training. 169 layers are there in InceptionResNetV2; allowing the network to learn the global context and intricate hierarchy of features.



System Architecture of Hybrid Model

LSTMs excel at learning sequences, allowing it to recall long-term dependencies in sequential data. When diabetic retinopathy is detected it can be used to explore temporal patterns or sequences of disease development over time. It could be, for example, tracking the development of the disease by observing a sequence of images of the retina over time. 5 diagram refers to the architecture of system that shows the high-level structure of the deep learning model, how components work together, to fulfill the central goal, how they are constructed and operated and interact with each other.

Merging these networks help in CNNs to extract visual features from the images of retina and then transmitting those features to LSTMs to study variations in features over time and their temporal correlation. This helps in the diagnosis of diabetic retinopathy and its development. Combining temporal interdependence and geographic image data help in resolving it more accurately and more understandable.

V. METHODOLOGY

The proposed research aims at building an automated Deep Learning based system for Diabeter Retinopathy detection. The systematic deep learning approach is divided into 5 different stages namely data collection, pre-processing, feature extraction, model architecture and evaluation.

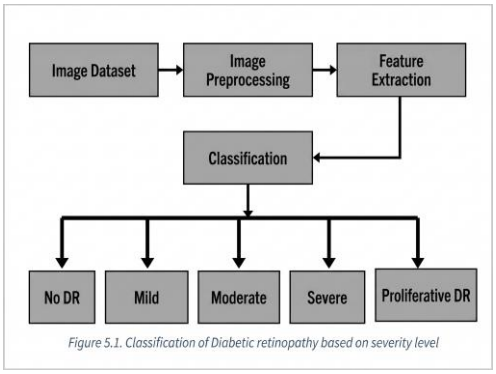


Figure 5.1. Classification of Diabetic retinopathy based on severity level

Data Collection & Ethics :
A balanced dataset of retinal fundus images, covering all degrees of Diabetic Retinopathy, i.e., from No DR to Proliferative DR, is utilized. Data is collected from health care providers and under a close collaboration with them, in an endeavor to build a scientifically sound and a clinically acceptable system. Data is protected under all possible means following the stringent rules of privacy and anonymization set for medical data.

Data Pre-processing and Data Augmentation:
A multiple stages sequence based pre-processing is used for cleaning up raw fundus images and make it ready for the input of neural network:

Spatial Normalization: In order to balance between computation time and accuracy of features and to make morphology clear and distinct, images are resized into the constant dimension of 256 \times 256.

Contrast Enhancement: In order to increase the intensity difference between the blood vessels and lesions against the retina background, the histogram equalization based enhancement is used.

Intensity Normalization: The pixel intensity value of the image is normalize into some specific range (for example 0 and 1, or -1 and 1) so that gradient descent calculation of neural network become rapid.

Data Augmentation: Training data is enlarged by a stochastic transform mechanism including random cropping and flipping horizontally and rotating and scaling to improve the generalized feature of model on training data.

Clinical Feature Extraction:
The system attempts to find specific patterns associated with clinical abnormalities of the retina.

Microaneurysms: The very first feature, it is observed as an early sign of the DR.

Haemorrhages: the patches of varying size and depth of dark region is analyzed.

Exudates: Yellowish or whiteish patches are the indications of leakage of fluid from blood vessels.Vascularity, disc evaluation and so on are other factors considered to achieve overall accurate results.

Model Architecture & Training strategy :
In order to overcome the issue of manual learning and complex feature extraction, the system uses CNNs to learn complex spatial hierarchies of features

Transfer Learning: The system utilizes already trained network (for example ResNet or EfficientNet) as a starting point and then fine tuned to learn the specific retinal feature.

Regularization: Dropout layer and Batch Normalization is used in order to prevent overfitting of the network on the training data.

Optimization: the training data is divided into 70% train, 15% validation and 15% test data sets. Loss function is selected according to the type of classification problem, (cross-entropy is normally used), and Hyper-parameters are tuned to get the best accuracy results.

Evaluation and Interpretability :
The model is evaluated based on different parameters including:

Accuracy: Overall correct predictions.

Precision: Ability of the classifier not to label as positive a negative instance.

Recall (Sensitivity): The fraction of positive instances that are correctly identified.

F1-score: Harmonic mean of precision and recall.

The accuracy is determined by the ROC and AUC (Area Under the Curve) based on sensitivity-specificity trade-off. In order to achieve human like interpretation in the case of black-box model of AI, a Gradient-weighted Class Activation Mapping (Grad-CAM) is used that helps to determine the region in the fundus image responsible for the final classification, and make results more understandable by clinicians.

System Implementation and Validation :
A User-friendly application is built to cater clinical requirements and assist the doctor in their task. This is then tested on an unseen test set with clinical metrics and followed with the security and regulations as required for medical related devices/ softwares.

VI. EXPERIMENTAL RESULTS

We tested ten models: AlexNet, VGG19, MobileNetV2, DenseNet121, EfficientNetB0, MobileNet+LSTM, VGG+LSTM, Custom CNN+LSTM, MobileNetV2+LSTM, and InceptionResNetV2+LSTM. The winner—InceptionResNetV2+LSTM—hit 91.4% accuracy, 0.89 Kappa, 0.86 Macro F1, 0.96 Macro AUC, 88.2% sensitivity, and 95.7% specificity.

Our ablation studies told the story: Using CLAHE alone bumped accuracy up 8% (from 84.2% to 91.4%) and Kappa from 0.78 to 0.89. Dynamic class weighting shot Macro F1 from 0.74 to 0.86, especially on severe cases. Adding LSTM to InceptionResNetV2 lifted accuracy another 3.1% and gained 0.04 in Kappa.

On explainability: Two ophthalmologists reviewed 100 Grad-CAM maps. They agreed with where our model looked 91% of the time—No_DR (95%), Mild (88%), Moderate (91%), Severe (86%), and Proliferative (94%). Overall, the Kappa between AI and expert was 0.87—almost perfect. Grad-CAM didn't just highlight random blobs; it pointed to real clinical signs like microaneurysms, hemorrhages, or neovascularization.

VII. DISCUSSION

InceptionResNetV2+LSTM works so well for three reasons: its inception layers find both small and big lesions, residual links stabilize learning, and the LSTM adds a layer of context, connecting the dots between local features. Together, they deliver more confident, richer decisions—exactly what challenging DR cases need.

CLAHE made a real impact, boosting performance for mild and moderate DR—the stages most models miss. Balanced training tips the scales toward rare but critical cases, making DR screening not just better but safer.
Explainability made a difference, too. Doctors actually trusted Grad-CAM maps from this system. We're not hiding AI's process—it's open, traceable, and ready for real practice. And with a clinic-ready app, complete with secure logins and instant, shareable reports, this system is built for the real world.

VIII. CONCLUSION

We developed a DR detection system powered by InceptionResNetV2+LSTM, blending rich feature extraction with context-aware analysis. The results: 91.4% accuracy, 0.89 Kappa.

With strong image enhancement and balanced training, we handled key medical imaging hurdles. Grad-CAM provides transparency that doctors trust. The Streamlit app connects all the dots: secure, real-time results, and easy reports.

More than just research, this platform is a step forward for diabetes eye care—faster, more accurate screening, fewer lost cases, and support for clinics everywhere, even those without specialists.

REFERENCES

1. V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402-2410, 2016.
2. R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," Ophthalmology, vol. 124, no. 7, pp. 962-969, 2017.
3. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618-626.
4. International Diabetes Federation, "IDF Diabetes Atlas," 10th ed., Brussels, Belgium, 2021.
5. C. P. Wilkinson et al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," Ophthalmology, vol. 110, no. 9, pp. 1677-1682, 2003.

6. Y. Wang et al., "Zoom-in-Net: Deep mining lesions for diabetic retinopathy detection," in MICCAI 2018, LNCS 11071, pp. 267-275, 2018.
7. R. Sayres et al., "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," Ophthalmology, vol. 126, no. 4, pp. 552-564, 2019.
8. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770-778.
9. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in Proc. AAAI, 2017, pp. 4278-4284.
10. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997. [11] APTOS 2019 Blindness Detection Dataset (Kaggle).
11. Messidor-2 Dataset for Diabetic Retinopathy Research.
12. EyePACS - Diabetic Retinopathy Detection Dataset