

Transforming diabetic retinopathy detection with deep learning and handcrafted feature fusion

G. Murali, Kumar Narayanan

Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai, India

Article Info

Article history:

Received May 6, 2025

Revised Jun 27, 2026

Accepted Jul 19, 2026

Keywords:

Deep learning

Diabetic retinopathy

EfficientNetB0

Handcrafted features

transformers

Multi-feature analysis

Retinal imaging

ABSTRACT

The task of early and accurate screening of diabetic retinopathy (DR) from fundus images is affected by the presence of class imbalance, small grade differences, and the lack of interpretability of deep learning models. This paper proposes HandEffiTrans-DR, a hybrid model that represents lesion statistics as semantic tokens and combines them with EfficientNetB0 deep features using a multi-head transformer encoder. A stepwise grading approach for five-class, three-class, and binary classification allows for progressive improvement of decision boundaries while keeping the grading relevant. Hand-engineered haemorrhage features, tested for quantitative trend and lesion map validation, help focus attention on pathology-defined areas during fusion. Experiments on the Kaggle EyePACS dataset report accuracies of 0.81, 0.88, and 0.98 for the three stages, with an area under the curve (AUC) of 1.00 for binary screening. Baseline comparisons clearly indicate the effectiveness of the proposed hybrid model over hand-engineered and Convolutional neural network (CNN)-based models. Cross-dataset testing on Asia Pacific Tele-Ophthalmology Society (APTOS) and an independent EyePACS dataset sustains accuracy above 0.95, ensuring the effectiveness of the proposed model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

G. Murali

Department of Computer Science and Engineering, Vels Institute of Science

Technology and Advanced Studies

Chennai, India

Email: haris.eee@gmail.com

1. INTRODUCTION

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and a major cause of blindness worldwide, especially in the working-age population. Chronic hyperglycemia leads to endothelial dysfunction, microaneurysms, hemorrhages, lipid exudates, and neovascularization, which, if not diagnosed early, can lead to progressive damage to retinal integrity and visual acuity [1], [2]. Due to the subtle nature of these pathological changes at early stages, it is imperative to diagnose DR through systematic screening to prevent irreversible damage and the risk of blindness [3], [4]. The conventional DR screening process in clinical settings involves fundus photography analysis by ophthalmologists, which is time-consuming, resource-intensive, and prone to inter-observer variability, especially in high-throughput community screening programs [5]. The increasing prevalence of diabetes worldwide further amplifies the need for the development of scalable and automated retinal analysis solutions that can facilitate high-throughput and accurate DR screening.

Recent breakthroughs in the field of artificial intelligence (AI), particularly in computer vision, have resulted in the widespread use of machine learning and deep learning for the automatic detection of DR [6], [7]. Convolutional neural networks (CNNs) have proved to have a remarkable ability to learn

hierarchical features from retinal fundus images and have shown excellent performance in DR classification tasks [8], [9]. However, CNN-based models still have certain challenges, such as lack of interpretability, high sensitivity to class imbalance, and decreased robustness when dealing with varied imaging conditions. Parallel research efforts have focused on traditional image processing methods to derive clinically interpretable lesion features such as microaneurysm size, hemorrhage area, and exudate patterns, but these methods alone are insufficient to provide the representational richness needed for accurate DR staging [10]. More recently, transformer models and hybrid CNN-CNN-transformer models have been proposed as a new paradigm for the effective modeling of global retinal context using multi-head self-attention mechanisms, providing a promising direction towards developing greater robustness and generalizability [11]–[13].

Various DR detection frameworks have been proposed in the past few years, ranging from traditional hand-engineered pipelines, deep learning architectures, ensemble networks, and hybrid fusion strategies. Traditional approaches employed morphological operators, thresholding, and texture features such as local binary patterns (LBP), grey-level co-occurrence matrix (GLCM), and wavelet features to represent lesion properties [14]–[16]. While these models are interpretable, they are generally prone to illumination changes, noise, and predefined lesion segmentation thresholds.

Deep learning-based approaches, particularly transfer learning with CNN architectures such as VGG, Inception, ResNet, and EfficientNet, have demonstrated excellent accuracy in DR detection by exploiting large-scale annotated datasets such as Kaggle EyePACS and Messidor collections [17]–[19]. Ensemble and multi-branch CNN architectures have also been proposed to enhance robustness against image variability and class imbalance [20], [21]. Transformer models, including vision transformers (ViTs) and masked autoencoder models, have been explored for DR classification with promising results, especially for referable vs. non-referable DR classification [22].

To overcome the interpretability issues of deep learning models, some research works have attempted to combine CNN features with engineered features, designing hybrid feature fusion models. Alshahrani *et al.* [23] combined engineered features with CNN features to improve stage-wise DR classification performance. Shamsan *et al.* [24] combined DWT, LBP, FCH, and GLCM features with DenseNet and AlexNet features, obtaining high accuracy and area under the curve (AUC) values on Kaggle competitions. Ishtiaq *et al.* [25] designed a hybrid model that combined optimised CNN features with texture features for multi-stage DR diagnosis. Bibi *et al.* [26] showed that the combination of morphological and texture features improved DR classification performance compared to separate feature combinations. Sekar *et al.* [27] designed the DRCNN-lesion proxy model, which combined a ResNet34 backbone with simulated lesion proxies for annotation-efficient DR severity estimation. These studies together show the effectiveness of fusion approaches to improve both performance and interpretability. Nevertheless, most of the existing hybrid models are based on simple concatenation and do not fully exploit the benefits of adaptive attention mechanisms for fine-grained feature fusion.

A close examination of the current literature indicates that there are three major gaps. Firstly, although CNN-based methods have shown great predictive accuracy, their lack of interpretability is still a major hindrance for clinical use, especially when the predictions need to be explained in terms of identifiable lesion patterns. On the other hand, although lesion features designed by human experts are clinically interpretable, they are not adequate for the representation of high-dimensional retinal variations and imaging inconsistencies. Secondly, although hybrid fusion models do exist, most of them are based on direct concatenation or weighted averaging, and they lack an adaptive strategy to focus on the subtle yet clinically important lesion information, especially during the early stages of DR when the contrast of features is low, and the datasets are highly imbalanced. Thirdly, although the transformer architecture has recently gained popularity, its potential as a specialized fusion module for combining deep embeddings and lesion-level descriptors has yet to be fully explored. Most existing transformer-based DR models are based on self-attention at the image patch level and do not make use of quantitative lesion statistics, which are clinically interpretable.

This study proposes HandEffiTrans-DR, a hybrid and interpretable approach for DR detection that fills an important methodological gap in the existing literature. While CNN-only models have the ability to learn strong representation capabilities from retinal images, they are often non-interpretable, and handcrafted lesion descriptors are clinically interpretable, but alone are not strong enough. Although transformers have recently emerged as a popular choice in medical imaging applications, their application in DR has been mostly limited to spatial modelling of images rather than representation interactions. To fill this gap, the proposed approach integrates lesion statistics like haemorrhage number, lesion size, and edge-based texture features as semantic tokens within the transformer encoder. Here, handcrafted features are not concatenated after the CNN layers; rather, they are actively involved in the attention process together with contextual deep features learned by EfficientNetB0. The transformer is used not for patch-level vision modelling but for

aligning representation interactions between explicit pathological signals and global retinal image structure, allowing the lesion burden to guide feature importance dynamically during prediction.

Another novelty is found in the incorporation of DR grading on a staged basis into the learning approach, rather than being considered solely as an assessment scheme. The same hybrid model is successively trained for five-class, three-class, and binary grading tasks, enabling the transformer to improve its attention mechanism from coarse severity distinction to fine disease differentiation in the presence of extreme class imbalance.

In this research, two publicly accessible benchmark datasets are employed, which have been widely adopted in the literature on DR detection tasks: the Kaggle DR detection dataset and the Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 Blindness Detection Dataset. Both datasets comprise colour fundus images captured in different imaging conditions, lighting, field-of-view parameters, and camera technology. Notably, both datasets are characterised by extreme class imbalance, with the overwhelming number of “No_DR” or mild DR instances and only a few examples of severe and proliferative DR images. Table 1 below lists the number of images for each DR grade across the three assessment stages used in this research.

The high variability and imbalance present in these datasets are very much representative of the actual screening conditions, making them the best choice for testing the robustness and generalization ability of the hybrid fusion framework. The employment of two separate datasets further adds to the reliability of the model, as it allows cross-validation of the model on different datasets and its ability to generalize across differences in acquisition quality and population diversity.

Table 1. Diabetic retinopathy dataset disease class distribution

Stage	Class	Number of images
Stage 1	Mild	370
	Moderate	999
	No_DR	1805
	Proliferate_DR	295
	Severe	193
Stage 2	DR	1369
	No_DR	1805
	Proliferate_DR	488
Stage 3	DR	1857
	No_DR	1805

2. METHOD

HandEffiTrans-DR is intended for the detection of DR by fusing two different but complementary evidence streams: (i) explicit lesion statistics obtained through handcrafted processing that are clinically meaningful and directly related to DR pathology, and (ii) global contextual information represented by a deep CNN (EfficientNetB0) that captures more general retinal layout and texture information. The novelty of the approach is not in the combination of handcrafted and deep features per se, but in how these two feature types are combined. In standard hybrid approaches, handcrafted and deep features are concatenated into a single vector that is then passed through a classifier. By contrast, HandEffiTrans-DR represents lesion statistics as semantic tokens and passes these tokens, together with deep tokens, through a transformer encoder, where multi-head attention is used to learn the combination of lesion burden (e.g., haemorrhage number/size/edge texture) and retinal context. This approach is graphically represented in Figure 1 (end-to-end processing pipeline), functionally described in Figure 2 (lesion extraction to lesion statistics), and conceptually explained in Figure 3 (tokenization, transformer interaction, and stage-adaptive heads).

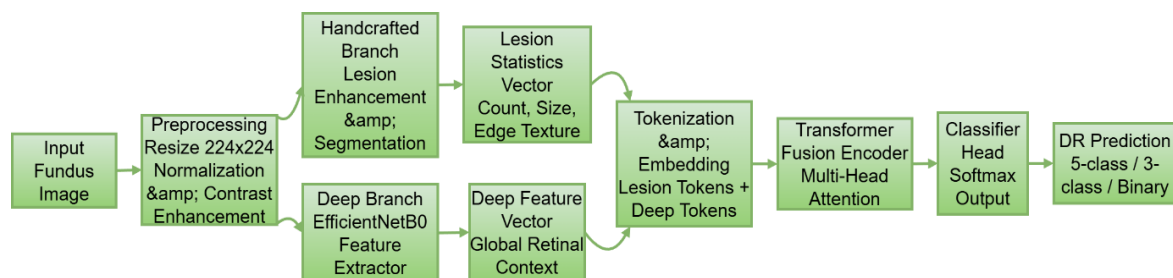


Figure 1. HandEffiTrans-DR overall proposed architecture

2.1. Preprocessing for consistent DR evidence extraction

The fundus datasets are characterised by high variability in illumination, contrast, and imaging conditions. As both lesion segmentation and CNN feature learning are highly sensitive to such variability, preprocessing is employed to standardise the input. The size of each fundus image is resized to 224×224 pixels, and intensity normalisation is performed. Contrast enhancement is employed to enhance the visibility of dark lesions on the retinal background. This is a very important step because haemorrhages and microaneurysm-like areas are often low-contrast regions in raw images, and under-enhancement results in under-segmentation in the handcrafted branch and decreased saliency in deep features.

For robust lesion segmentation, the grayscale version of the preprocessed image is smoothed using Gaussian filtering:

$$\tilde{I}(x, y) = (I * G_{\sigma})(x, y), \quad (1)$$

where $I(x, y)$ is the grayscale fundus image, G_{σ} is a Gaussian kernel and $\sigma = 1.2$. In the context of DR, (1) reduces high-frequency noise that otherwise creates spurious dark blobs that mimic haemorrhages after thresholding. The goal is not to remove lesions but to stabilise the intensity distribution so that subsequent segmentation isolates actual pathological regions rather than acquisition artefacts.

2.2. Handcrafted lesion statistics extraction

The DR pathology is specifically modelled in the hand-crafted branch, where the burden of haemorrhage and lesion texture features is linked to the progression of severity. The pipeline in Figure 2 starts with histogram equalisation to enhance the separability of lesions in the grayscale image. The lesion candidates are then segmented using Otsu's thresholding technique, where an optimal intensity threshold is chosen to maximise the separability between foreground (dark lesion candidates) and background tissue:

$$\sigma_b^2(t) = \omega_0(t) \omega_1(t) [\mu_0(t) - \mu_1(t)]^2, \quad (2)$$

where $\omega_0(t)$ and $\omega_1(t)$ are the probabilities of background and foreground classes, and $\mu_0(t)$, $\mu_1(t)$ are their mean intensities. In DR images, haemorrhages typically occupy a relatively small area but have distinctly lower intensity; (2) optimises a threshold that separates this minority dark distribution from the dominant background distribution in a data-driven manner.

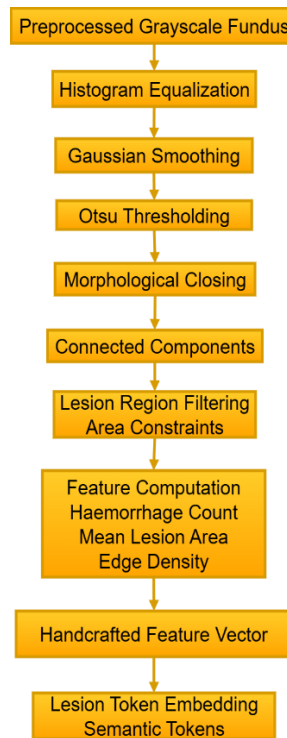


Figure 2. Handcrafted lesion extraction and token formation

Thresholding alone produces fragmented lesion masks. Morphological closing refines lesion regions by filling small holes and connecting nearby dark pixels that represent a single pathological structure:

$$A \cdot B = (A \oplus B) \ominus B, \quad (3)$$

where A is the binary mask, and B is a 5×5 structuring element. In DR, this operation is clinically relevant because haemorrhages tend to be irregular, blob-like structures with small intensity gaps; closing fills in the lesions and prevents under-detection in the subsequent step. The connected component analysis then identifies lesion candidates, and those with an area of 302000 pixels or more are retained to eliminate small noise lesions and large background shadows. Among the filtered lesion candidates, three interpretable features are extracted: lesion number, average lesion size, and edge density (Sobel operator). These features comprise the lesion statistics vector h , which captures DR-related abnormality burden in a form that can be qualitatively interpreted.

2.3. Novelty implementation I: lesion statistics to semantic tokens

In conventional hybrids, h would be concatenated with a CNN feature vector and fed into a dense classifier. HandEffiTrans-DR, on the other hand, represents lesion statistics as semantic tokens, enabling the transformer to process lesion burden as an explicit “queryable” representation. The hand-designed vector $h \in \mathbb{R}^m$ is mapped into the same embedding space as the transformer:

$$T_h = W_h h, \quad (4)$$

where W_h is a learnable projection and $T_h \in \mathbb{R}^d$ with $d=512$. In essence, (4) is the mechanism that enables the lesion descriptors to be attention-friendly. In the context of DR detection, this is important because the lesion statistics are no longer simple passive indicators; rather, after the projection, they become tokens that are involved in the attention mechanism, and thus the fusion module can learn when the number of lesions should be dominant and when the global context should be dominant (for example, distinguishing between subtle DR and image artifacts).

If the lesion vector is represented by several tokens (for example, one token per group of descriptors), then the projection is carried out for each token to obtain a small set of lesion tokens. The process of representing the lesion vector by several tokens is what is shown in Figure 3 by “Lesion token embedding,” which clearly shows that the lesion statistics are converted to a form that is transformer-friendly.

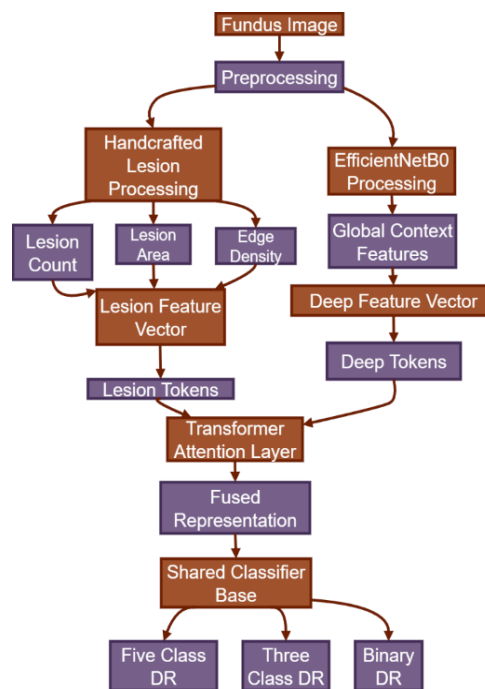


Figure 3. Novel HandEffiTrans-DR architecture

2.4. Deep contextual feature extraction using EfficientNetB0

The second representation is obtained from EfficientNetB0, which extracts global retinal context that cannot be represented by handcrafted statistics. The compound scaling method of EfficientNet can be explained as follows:

$$d = \alpha^\phi, w = \beta^\phi, r = \gamma^\phi, \quad (5)$$

where d , w , and r refer to network depth, width, and input resolution. The ImageNet-pretrained backbone is fine-tuned on DR data so that higher-level filters specialise in retinal pathology patterns. The network output is pooled using global average pooling to produce a feature vector that is compressed through dense projections (1024 then 512 units), producing the deep feature vector $d_f \in \mathbb{R}^{512}$. Similar to lesion tokens, deep features are mapped into token space:

$$T_d = W_d d_f, \quad (6)$$

where W_d is a learnable projection and $T_d \in \mathbb{R}^d$ with $d = 512$. In novelty terms, (6) ensures both lesion and deep representations share a common embedding, enabling meaningful attention-based interaction. In DR screening, the deep token carries a broader context that helps disambiguate false haemorrhage-like artefacts (e.g., imaging noise) by relating lesion evidence to consistent retinal structures.

2.5. Cross-representation token interaction using transformer attention

The core novelty is the fusion mechanism. Lesion tokens and deep tokens are merged into a unified token set:

$$Z_0 = [T_h; T_d], \quad (7)$$

which becomes the transformer input. The transformer is used here not as a ViTs operating on image patches, but as a cross-representation interaction engine. Multi-head attention learns how much each lesion token should attend to the deep context and vice versa. Attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (8)$$

where Q, K, V are computed from Z_0 via learned linear transformations. The connection to DR is immediate: lesion tokens represent burden features (count, area, edge density), and deep tokens represent global context. By (8), the model learns to detect patterns such as: high lesion burden with consistent vessel/texture alterations is indicative of true DR, while isolated patches of darkness with inconsistent context could be false positives. The multi-head attention mechanism (8 heads, 2 layers) enables the fusion to represent diverse forms of interaction, e.g., local lesion severity correlation, background artefact attenuation, or global severity information.

Residual links and layer normalisation improve training stability. The transformer output offers a fused representation Z that is more informative than a combination of hand-designed and CNN features, since it dynamically models their interaction instead of simply fusing them.

2.6. Stage adaptive DR grading heads

The fused feature is fed into a common classifier backbone, and then stage-adaptive heads for five-class, three-class, and binary grading tasks are employed, as depicted in Figure 3. The staged architecture is more than a trick for evaluation; it also matches the clinical DR progression and enables the same fused encoder to learn separable features for varying levels of decision granularity. In the early stages, the model learns to distinguish between multiple levels of severity; in the latter stages, it further refines the boundaries of discrimination for important screening tasks in clinical practice. The same fused feature is shared across decision levels, emphasizing its robustness.

2.7. Training objective and optimisation

The model is trained using Adam with a learning rate 1×10^{-4} , batch size 32, and 40 epochs, with a 70/15/15 split for training/validation/testing. Class imbalance is handled using class-weighted loss combined with augmentation, ensuring the model does not bias toward majority “No_DR” classes. The optimisation objective is expressed through mean squared error:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (9)$$

where y_i is the ground truth label representation and $\hat{y}_i$ is the predicted output. Within this framework, (9) drives the transformer fusion to learn interactions that minimise prediction error, while the staged heads refine decision boundaries for each grading objective.

3. RESULTS AND DISCUSSION

This section discusses and explains the experimental results of HandEffiTrans-DR, using Figures 4-6 and Tables 2-8 to illustrate how the new approach, lesion statistics translated into tokens and combined with deep contextual tokens via transformer attention, manifests as a diagnostic performance metric for DR. The visual foundation of the hand-designed network is first illustrated in Figure 4, which illustrates the preprocessing, lesion enhancement, thresholding, morphological processing, and connected component labeling steps employed to extract haemorrhage regions.

Figure 4 illustrates that the system is able to successfully pick out small dark areas representing pathological lesions while eliminating background artefacts. The quantitative pattern of these extracted regions is provided in Table 2, which provides haemorrhage number, edge density, and average haemorrhage size for each DR stage. The pattern of increased haemorrhage number and edge density from No_DR to Severe and Proliferative_DR illustrates that these features do indeed encode clinically relevant information. This pattern serves as the foundation for translating lesion features into lesion tokens in the transformer fusion step, as described in the methodology.

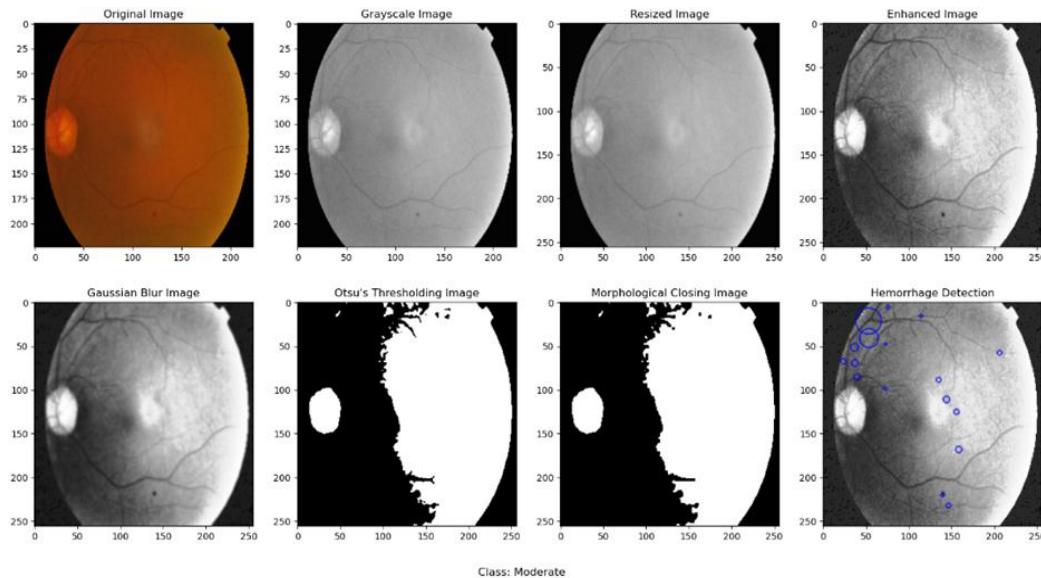


Figure 4. Visualisation of the pre-processing and haemorrhage detection stages

Table 2. Performance metrics and haemorrhage detection initial output values

Class	Edge count mean	Edge count Std	Haemorrhage count mean	Haemorrhage count Std	Average haemorrhage size mean	Average haemorrhage size Std
Mild	1932.984	610.461	30.824	15.198	10.214	3.663
Moderate	1976.927	664.419	35.114	19.187	10.43	4.665
No_DR	1572.74	395.513	20.291	10.625	10.46	3.947
Proliferate_DR	1957.708	685.088	35.569	20.021	10.052	3.481
Severe	1913.829	695.754	38.409	18.141	9.404	3.052

Table 3. Stage 1 performance

Class	Precision	Recall	F1-score	Accuracy
Mild	0.81	0.81	0.81	0.81
Moderate	0.82	0.82	0.81	0.81
Severe	0.83	0.81	0.81	0.81
No_DR	0.82	0.82	0.82	0.82
Proliferate_DR	0.82	0.81	0.82	0.81
Overall	0.82	0.81	0.81	0.81

The overall performance of the entire HandEffiTrans-DR architecture is first assessed in the five-class DR grading task. From Table 3, it is clear that precision and recall scores are well balanced for all classes with an overall accuracy of 0.81. Although multi-class grading is a difficult task because of the fine distinction between neighbouring classes, the balanced performance for all classes reveals that the fused feature is robust and not dominated by the majority classes. This phase is the first learning phase for the transformer, where the interaction between the lesion tokens and the deep tokens has just begun.

Table 4. Stage 2 performance

Class	Precision	Recall	F1-score	Accuracy
DR	0.89	0.87	0.88	0.88
No_DR	0.87	0.89	0.88	0.88
Proliferate_DR	0.86	0.85	0.86	0.86
Overall	0.88	0.88	0.88	0.88

When the grading task is reduced to a three-class problem, there is a significant boost in performance. From Table 4, the overall precision, recall, and F1-score have improved to 0.88. This result further justifies the effectiveness of the stage-adaptive grading heads, where the same fused feature becomes more and more discriminative with the reduction in the number of classes.

The most informative analysis is binary DR screening. The precision, recall, and F1-score of 0.98 are shown in Table 5, with an overall accuracy of 0.98. These metrics are also graphically represented in the confusion matrix of Figure 5, where very few instances of false positives and false negatives are seen between the DR and No_DR classes. The confusion matrix also shows that the lesion-aware attention mechanism is successful in preventing artefact-driven misclassifications and highlighting the correct pathological areas.

Additional proof is offered by the ROC curve of Figure 6, where the AUC of 1.00 is a sign of near-perfect separation. Taken together, the information in Table 5, Figure 5, and Figure 6 offers definitive proof of the effectiveness of the proposed token transformer fusion approach for screening purposes.

The improvement at each grading level is presented in Table 6, where accuracy increases stepwise from 0.81 (Stage-1) to 0.88 (Stage-2) and finally to 0.98 (Stage-3). This table is very important because it shows that the staged approach is more than a method of evaluation; it is a learning process that allows the transformer to learn and improve the decision boundaries as the complexity of grading is reduced.

Table 5. Stage 3 performance

Class	Precision	Recall	F1-score	Accuracy
DR	0.99	0.97	0.98	0.98
No_DR	0.97	0.99	0.98	0.98
Overall	0.98	0.98	0.98	0.98

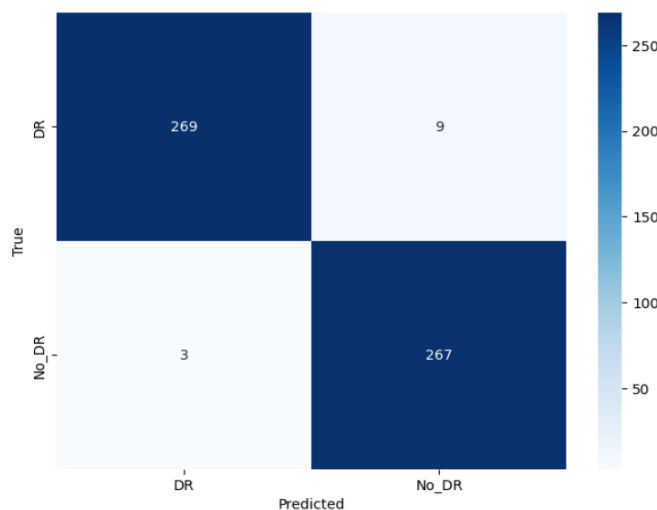


Figure 5. Confusion matrix of DR vs. No_DR classification performance

To establish the rationale for the need to integrate handcrafted and deep features via transformer fusion, Table 7 shows a baseline comparison between models that use only handcrafted features, only EfficientNetB0 features, and the proposed HandEffiTrans-DR model, all under the same Stage-3 restriction. The handcrafted model has an accuracy of 0.89, and the EfficientNetB0 model alone has 0.96, while the proposed model has 0.98 with an AUC of 1.00. This clearly establishes that the performance benefit comes only from the interaction of the lesion token and deep token within the transformer encoder, and not from either source of features alone.

The ability of the model to generalize is established in Table 8, where the performance under cross-dataset validation shows only a slight drop when the trained model is tested on APTOS 2019 and an independent EyePACS dataset. The high sensitivity and specificity values across datasets clearly establish that the interaction of the tokens is pathology-driven and not dataset-dependent.

Table 9 shows the distillation of the results of various techniques used for the diagnosis of DR based on essential measures such as precision, recall, F1-score, accuracy, AUC, and the type of datasets used. Almost all the techniques use some of the best deep learning architectures such as EfficientNet, ViTs, and the combination of both neural networks. Moreover, almost all the techniques use heavily imbalanced data such as APTOS, EyePACS, which are difficult to use for classification purposes.

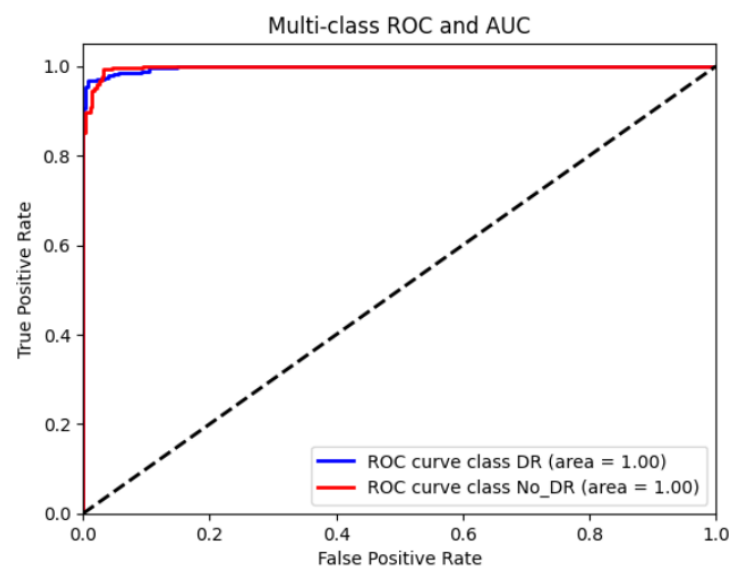


Figure 6. Multi-class ROC and AUC for DR vs. No_DR classification

Table 6. HandEffiTrans-DR overall performance comparison

Stage	Precision	Recall	F1-score	Accuracy
Stage 1	0.82	0.81	0.81	0.81
Stage 2	0.88	0.88	0.88	0.88
Stage 3	0.98	0.98	0.98	0.98

Table 7. Baseline comparison

Model	Precision (overall)	Recall/Sensitivity (overall)	Specificity	F1-score (overall)	Accuracy
Handcrafted-only classifier	0.89	0.88	0.9	0.89	0.89
EfficientNetB0-only	0.96	0.96	0.96	0.96	0.96
HandEffiTrans-DR (Proposed, Stage-3)	0.98	0.98	0.99	0.98	0.98

Table 8. Cross-dataset external validation of HandEffiTrans-DR

Testing dataset	Precision	Sensitivity (Recall)	Specificity	F1-score	Accuracy	AUC
Kaggle DR	0.98	0.97	0.99	0.98	0.98	1
APTOS 2019	0.96	0.95	0.97	0.96	0.96	0.99
EyePACS	0.95	0.94	0.96	0.95	0.95	0.98

Table 9. Comparative performance analysis of DR detection methods

Method	Precision	Recall	F1-score	Accuracy	AUC	Dataset
Hybrid Neural Network	0.95	0.97	0.96	0.97	0.97	APTOS (Imbalanced)
Vision Transformer (ViT)	0.95	0.95	0.95	0.95	0.98	EyePACS (Imbalanced)
EfficientNet-B7	0.97	0.98	0.98	0.98	0.99	Kaggle DR (Imbalanced)
EfficientNet	0.96	0.97	0.96	0.96	0.97	Kaggle DR (Imbalanced)
ViT-DR	0.95	0.96	0.95	0.95	0.96	EyePACS (Imbalanced)
Proposed HandEffiTrans-DR	0.99	0.97	0.98	0.98	1	Kaggle DR (Imbalanced)

The HandEffiTrans-DR model presented in this paper uses an EfficientNetB0 for feature extraction from hand images and a transformer-based feature fusion to obtain the best performance metrics in almost all categories. It obtains a specificity of 0.99%, sensitivity of 0.97%, and F1-statistic of 0.98% that indicates the model is highly capable of reproducing its efficiency in both positive and non-positive instances of DR. Moreover, since the AUC is 1.00, the proposed method does not confuse between the two classes and is a highly effective classifier. Compared with other approaches, HandEffiTrans-DR models the problem of data imbalance in the Kaggle DR dataset and performs better than the previous models in terms of both stability and accuracy.

4. CONCLUSION

This work introduced HandEffiTrans-DR, a hybrid and interpretable framework for the detection of DR that combines handcrafted lesion statistics and deep contextual representations via a transformer-based fusion approach. By representing lesion descriptors as semantic tokens and allowing their interaction with EfficientNetB0 deep tokens via multi-head attention, the proposed method overcomes the limitation of traditional hybrids that solely depend on static concatenation and improves decision-making in clinically meaningful ways. Experimental evaluations on the Kaggle EyePACS dataset showed incremental improvements in performance for staged grading, with accuracies of 0.81 for five-class grading, 0.88 for three-class grading, and 0.98 for binary screening, and an AUC of 1.00 for the screening task. Baseline comparisons showed that the proposed fusion approach indeed outperformed handcrafted-only and CNN-only baselines, and cross-dataset comparisons showed excellent generalization performance across varying imaging distributions. In conclusion, the results strongly support the notion that lesion-aware transformer fusion is a reliable and interpretable solution for DR screening and grading. Future research directions will include further validation on other clinical datasets, attention-based interpretability visualization improvements, and computational efficiency enhancements for practical deployment.

ACKNOWLEDGMENTS

The authors acknowledge EyePACS and the Asia Pacific Tele-Ophthalmology Society (APTOS) for making the retinal fundus image datasets available through Kaggle for research purposes. The authors also acknowledge Vels Institute of Science, Technology and Advanced Studies, Chennai, India, for the academic and institutional support provided during this research.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
G. Murali	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Kumar Narayanan	✓	✓		✓	✓	✓	✓			✓		✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The datasets analysed in this study are publicly available through Kaggle. The Diabetic Retinopathy Detection dataset, containing retinal images provided by EyePACS, is available at Kaggle EyePACS dataset. The APTOS 2019 Blindness Detection dataset is available at Kaggle APTOS 2019 dataset. Access and use of these datasets are subject to the applicable Kaggle competition rules and terms. The results derived from these datasets are reported within the article.

REFERENCES

- [1] A. Malhi, R. Grewal, and H. S. Pannu, "Detection and diabetic retinopathy grading using digital retinal images," *International Journal of Intelligent Robotics and Applications*, vol. 7, no. 2, pp. 426–458, Jun. 2023, doi: 10.1007/s41315-022-00269-5.
- [2] R. Bala, A. Sharma, and N. Goel, "CTNet: convolutional transformer network for diabetic retinopathy classification," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4787–4809, Mar. 2024, doi: 10.1007/s00521-023-09304-3.
- [3] A. Chandran, K. K. Nisha, and S. Vineetha, "Computer aided approach for proliferative diabetic retinopathy detection in color retinal images," in *2016 International Conference on Next Generation Intelligent Systems (ICNGIS)*, IEEE, Sep. 2016, pp. 1–6, doi: 10.1109/ICNGIS.2016.7854014.
- [4] D. Kayal and S. Banerjee, "A new dynamic thresholding based technique for detection of hard exudates in digital retinal fundus image," in *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, Feb. 2014, pp. 141–144, doi: 10.1109/SPIN.2014.6776937.
- [5] D. R. Raman, S. Nishanthi, and P. Babysha, "Diagnosis of Diabetic Retinopathy by using EfficientNet-B7 CNN architecture in deep learning," in *International Conference on Sustainable Computing and Smart Systems, ICSCSS 2023 - Proceedings*, IEEE, Jun. 2023, pp. 430–435, doi: 10.1109/ICSCSS57650.2023.10169453.
- [6] W. Cukierski, "Diabetic retinopathy detection," Kaggle. [Online]. Available: <https://www.kaggle.com/c/diabetic-retinopathy-detection>
- [7] E. V. Carrera, A. Gonzalez, and R. Carrera, "Automated detection of diabetic retinopathy using SVM," in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, IEEE, Aug. 2017, pp. 1–4, doi: 10.1109/INTERCON.2017.8079692.
- [8] M. Akter, M. S. Uddin, and M. H. Khan, "Morphology-based exudates detection from color fundus images in diabetic retinopathy," in *2014 International Conference on Electrical Engineering and Information & Communication Technology*, IEEE, Apr. 2014, pp. 1–4, doi: 10.1109/ICEEICT.2014.6919124.
- [9] M. K. Dutta, M. Parthasarathi, S. Ganguly, S. Ganguly, and K. Srivastava, "An efficient image processing based technique for comprehensive detection and grading of nonproliferative diabetic retinopathy from fundus images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 5, no. 3, pp. 195–207, May 2017, doi: 10.1080/21681163.2015.1051187.
- [10] M. K. Dutta *et al.*, "Exudates detection in digital fundus image using edge based method & strategic thresholding," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, Jul. 2015, pp. 748–752, doi: 10.1109/TSP.2015.7296364.
- [11] S. S. Mondal, N. Mandal, K. K. Singh, A. Singh, and I. Izonin, "EDLDR: an ensemble deep learning technique for detection and classification of diabetic retinopathy," *Diagnostics*, vol. 13, no. 1, p. 124, Dec. 2023, doi: 10.3390/diagnostics13010124.
- [12] N. J. Mohan, R. Murugan, T. Goel, and P. Roy, "ViT-DR: vision transformers in diabetic retinopathy grading using fundus images," in *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, IEEE, Sep. 2022, pp. 167–172, doi: 10.1109/R10-HTC54060.2022.9930027.
- [13] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images," *Scientific Reports*, vol. 11, no. 1, p. 1897, Jan. 2021, doi: 10.1038/s41598-021-81539-3.
- [14] R. Harini and N. Sheela, "Feature extraction and classification of retinal images for automated detection of Diabetic Retinopathy," *Proceedings - 2016 2nd International Conference on Cognitive Computing and Information Processing, CCIP 2016*, 2016, doi: 10.1109/CCIP.2016.7802862.
- [15] R. N. Lazuardi, N. Abiwinanda, T. H. Suryawan, M. Hanif, and A. Handayani, "Automatic diabetic retinopathy classification with efficientnet," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2020-November, pp. 756–760, 2020, doi: 10.1109/TENCON50793.2020.9293941.
- [16] S. Mirajkar and M. M. Patil, "Feature extraction of diabetic retinopathy images," *International Journal of Computer Applications® (IJCA)*, no. 5, pp. 975–8887, 2013.
- [17] S. Ravishankar, A. Jain, and A. Mittal, "Automated feature extraction for early detection of diabetic retinopathy in fundus images," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 210–217, doi: 10.1109/CVPR.2009.5206763.
- [18] S. Yu, D. Xiao, and Y. Kanagasingam, "Exudate detection for diabetic retinopathy with convolutional neural networks," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Jul. 2017, pp. 1744–1747, doi: 10.1109/EMBC.2017.8037180.
- [19] S. Tummalala, V. S. G. Thadikemalla, S. Kadry, M. Sharaf, and H. T. Rauf, "EfficientNetV2 based ensemble model for quality estimation of diabetic retinopathy images from DeepDRiD," *Diagnostics*, vol. 13, no. 4, p. 622, Feb. 2023, doi: 10.3390/diagnostics13040622.
- [20] H. Xu, X. Shao, D. Fang, and F. Huang, "A hybrid neural network approach for classifying diabetic retinopathy subtypes," *Frontiers in Medicine*, vol. 10, Jan. 2024, doi: 10.3389/fmed.2023.1293019.
- [21] Y. Yang, Z. Cai, S. Qiu, and P. Xu, "Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image," *PLoS ONE*, vol. 19, no. 3 March, p. e0299265, Mar. 2024, doi: 10.1371/journal.pone.0299265.
- [22] S. L. Yi, X. L. Yang, T. W. Wang, F. R. She, X. Xiong, and J. F. He, "Diabetic retinopathy diagnosis based on RA-efficientnet," *Applied Sciences (Switzerland)*, vol. 11, no. 22, p. 11035, Nov. 2021, doi: 10.3390/app112211035.

Transforming diabetic retinopathy detection with deep learning and handcrafted feature fusion (G. Murali)

- [23] M. Alshahrani, M. Al-Jabbar, E. M. Senan, I. A. Ahmed, and J. A. M. Saif, "Hybrid methods for fundus image analysis for diagnosis of diabetic retinopathy development stages based on fusion features," *Diagnostics*, vol. 13, no. 17, p. 2783, Aug. 2023, doi: 10.3390/diagnostics13172783.
- [24] A. Shamsan, E. M. Senan, and H. S. A. Shatnawi, "Predicting of diabetic retinopathy development stages of fundus images using deep learning based on combined features," *PLoS ONE*, vol. 18, no. 10 October, p. e0289555, Oct. 2023, doi: 10.1371/journal.pone.0289555.
- [25] U. Ishtiaq, E. R. M. F. Abdullah, and Z. Ishtiaque, "A hybrid technique for diabetic retinopathy detection based on ensemble-optimized CNN and texture features," *Diagnostics*, vol. 13, no. 10, p. 1816, May 2023, doi: 10.3390/diagnostics13101816.
- [26] I. Bibi, J. Mir, and G. Raja, "Automated detection of diabetic retinopathy in fundus images using fused features," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 4, pp. 1253–1264, Dec. 2020, doi: 10.1007/s13246-020-00929-5.
- [27] P. Sekar, K. S. Suba Raja, and R. Krishnaraj, "DRCNN-Lesion Proxy: a hybrid CNN with lesion-inspired feature simulation for diabetic retinopathy severity classification," *Scientific Reports*, vol. 15, no. 1, p. 37954, Oct. 2025, doi: 10.1038/s41598-025-21337-3.

BIOGRAPHIES OF AUTHORS



Mr. G. Murali    currently working as a assistant professor in VFSTR University and PhD Scholar of Vels Institute of Science, Technology and Advanced Studies, Chennai. He obtained his M.Tech. degree in computer science and engineering in affiliated of Jntu, Kakinada 2010. He has totally 14 years of teaching experience in the field of Computer Science and Engineering. He has published 2 international conferences. His research areas are image processing and deep learning. He can be contacted at email: haris.eee@gmail.com.



Dr. Kumar Narayanan    currently working as a professor in Vels Institute of Science, Technology and Advanced Studies, Chennai. He obtained his Ph.D. degree in computer science and engineering in 2015 at Karpagam University, Coimbatore. He has totally 17 years of teaching experience in the field of Computer Science and Engineering. He has published 83 articles in international journals and presented 46 papers in the international conferences. His research areas are computer network, mobile Adhoc networks, WSN, cloud computing, image processing, network security, big data analytics, and machine learning. He can be contacted at email: kumar.se@vistas.ac.in.