International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 3, September 2026, pp. 1112~1122
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i3.pp1112-1122

Lesion-semantic token guided transformer fusion for stage-adaptive diabetic retinopathy grading and screening

Murali Gujjula, Kumar Narayanan
Department of Computer Science and Engineering, School of Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai, India

Article Info

Article history:

Received May 6, 2025
Revised Aug 2, 2026
Accepted Aug 13, 2026

Keywords:

Diabetic retinopathy
Lesion-aware tokenization
Probability calibration
Retinal fundus imaging
Transformer attention

ABSTRACT

Diabetic retinopathy (DR) screening from retinal fundus images is still difficult because of the heterogeneity of lesions, significant class imbalance, and low reliability of existing deep learning algorithms owing to their generic feature pooling and passive fusion mechanisms. To overcome such limitations, this study presents a hybrid model called HandEffiTrans-DR, which combines deep retinal representations with handcrafted lesion descriptors via the directed transformer-based attention mechanism. First, a lesion-prior-guided token squeezer (LPG-TS) module is proposed in the work, which is capable of creating lesion-specific visual tokens via integrating prior knowledge about pathology into the tokenization procedure. Moreover, a directed transformer-based fusion (T→V) strategy allows for actively fusing the handcrafted lesion features with visual representations. The introduced framework is assessed with a stratified five-fold cross-validation protocol. Its average classification accuracy surpasses 97%, while the macro-F1 score is higher than 0.97 and Cohen's kappa is greater than 0.95, indicating good agreement despite the presence of class imbalance. The probability reliability is measured in terms of expected calibration error (ECE) and Brier scores and shows well-calibrated confidence. Further validation is performed on two external datasets, where consistent performance with minor decline is shown.

Corresponding Author:

Murali Gujjula
Department of Computer Science and Engineering, School of 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 remains one of the leading causes of irreversible vision impairment worldwide. The disease manifests through retinal lesions such as microaneurysms, hemorrhages, hard exudates, and neovascularization, which often appear at early stages with subtle visual contrast and localized spatial distribution. Large-scale screening using fundus photography is therefore essential for early diagnosis and timely intervention, yet manual screening is limited by specialist availability, inter-observer variability, and increasing patient burden. As a result, automated DR detection using artificial intelligence has become an active research area, aiming to support ophthalmologists by providing fast, consistent, and scalable screening solutions [1], [2].

Deep learning, particularly convolutional neural networks (CNNs), has demonstrated strong performance in DR detection by learning hierarchical representations directly from retinal images [3], [4]. CNN-based systems have achieved high accuracy on public benchmarks and are increasingly explored for

real-world deployment [5]. However, despite their success, CNN-only approaches often function as black-box models, offering limited interpretability and reduced robustness when exposed to variations in acquisition devices, illumination conditions, or patient demographics [6], [7]. These limitations are critical in medical screening, where model reliability and transparency are as important as raw classification performance.

To overcome shortcomings of purely deep models, recent studies have explored hybrid frameworks that combine deep features with handcrafted retinal descriptors and attention mechanisms [8], [9]. Handcrafted features, such as texture statistics and lesion-level descriptors, encode clinically meaningful information that complements deep representations and enhances interpretability [10]. At the same time, transformer-based architectures and vision transformers have been introduced into retinal image analysis to model long-range dependencies and global context through self-attention mechanisms [11], [12]. These approaches have shown promise in capturing spatial relationships between lesions and anatomical structures that are difficult to model using local convolution alone [13].

More recent works have proposed CNN-transformer fusion models for DR detection and grading, reporting improved discrimination performance compared to standalone CNNs [14], [15]. Nevertheless, most existing fusion strategies rely on simple feature concatenation or symmetric attention, which treats all spatial regions uniformly and fails to explicitly emphasize lesion-dominant areas [16], [17]. Furthermore, handcrafted features, when incorporated, are often appended only at the classifier stage, limiting their influence on intermediate representation learning [18]. Another critical limitation is that many studies focus primarily on accuracy or area under the receiver operating characteristic curve (AUROC), while overlooking probability calibration and confidence reliability—factors that are crucial for clinical decision-making [19], [20].

Despite significant progress, several gaps remain in current DR detection methodologies. First, existing models rarely integrate explicit lesion priors into the tokenization or attention process, leading to diluted lesion information when features are globally pooled or uniformly partitioned. Second, the interaction between handcrafted lesion descriptors and deep visual representations is often weak or passive, preventing clinically meaningful guidance during feature fusion. Third, transformer-based fusion mechanisms lack directionality that reflects clinical reasoning, where lesion statistics should inform visual interpretation rather than the reverse [21], [22]. Finally, many reported systems are evaluated on imbalanced datasets using accuracy-centric metrics, without sufficient analysis of class-wise behavior, robustness across folds, or probability calibration, which limits their suitability for real-world screening scenarios [23], [24].

The problem addressed in this work is therefore the development of a DR detection framework that i) explicitly emphasizes lesion-relevant regions during representation learning; ii) enables active and interpretable interaction between handcrafted lesion features and deep visual tokens; and iii) remains robust and reliable under class imbalance and dataset variability. To address the identified gaps, this work proposes HandEffiTrans-DR, a lesion-aware hybrid deep learning framework for automated DR detection.

The proposed method introduces two key innovations. First, a lesion-prior-guided token squeezer (LPG-TS) generates lesion-centric visual tokens by explicitly incorporating lesion prior maps derived from retinal pre-processing. This ensures that token representations are spatially anchored to pathological regions rather than uniformly sampled background areas. Second, a directed transformer-based fusion mechanism (T→V) is employed, where handcrafted lesion descriptors guide the attention refinement of visual tokens, enabling clinically meaningful cross-modal reasoning. In addition, the framework integrates a lesion summary token to preserve global retinal context and employs calibration-aware evaluation to assess the reliability of predicted probabilities. The overall architecture is designed to balance discriminative performance, interpretability, and robustness, aligning with the practical requirements of DR screening systems.

2. METHOD

2.1. Dataset characteristics and class imbalance consideration

Public DR datasets are inherently class-imbalanced, with normal or mild cases significantly outnumbering moderate-to-severe disease samples. This imbalance poses a major challenge, as models trained without appropriate consideration may achieve high accuracy while underperforming on clinically critical minority classes. In this study, stratified five-fold cross-validation is employed to preserve class distribution across folds, ensuring fair and unbiased evaluation.

Moreover, performance is assessed using macro-averaged metrics, Cohen's kappa, and calibration measures, which provide a more informative assessment under imbalance compared to accuracy alone [25]. The proposed framework is evaluated on a primary DR dataset and further validated on independent datasets to assess generalization across different data distributions. By combining lesion-aware modeling, directed fusion, and imbalance-sensitive evaluation, the proposed approach aims to deliver a reliable and clinically meaningful solution for automated DR screening.

2.2. Overview of the proposed framework

The proposed framework, referred to as HandEffiTrans-DR, is a lesion-aware hybrid learning architecture designed for accurate and reliable DR detection from retinal fundus images. The complete processing pipeline is illustrated in Figure 1, which presents the end-to-end flow from image pre-processing to calibrated DR prediction. The central design philosophy of the framework is to ensure that pathological lesion information is explicitly embedded into both feature representation and fusion stages, rather than being implicitly learned through deep networks alone. To achieve this, the framework integrates lesion-centric visual tokenization, handcrafted retinal descriptors, and a directed transformer-based fusion mechanism. As shown in Figure 1, the methodology consists of five sequential stages: image pre-processing and lesion enhancement, parallel deep and handcrafted feature extraction, lesion-prior-guided token generation, directed transformer fusion, and calibrated classification. These stages are tightly coupled to maintain consistency between pathological evidence and final decision-making.

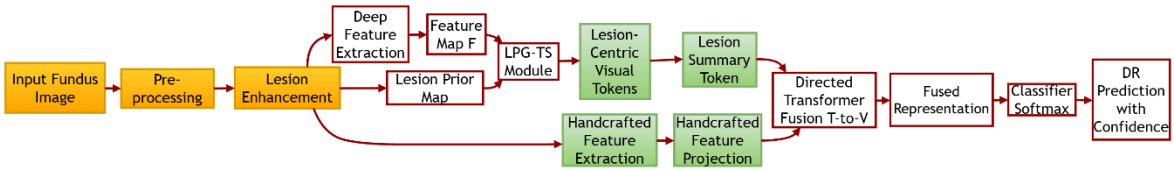


Figure 1. Overall HandEffiTrans-DR framework architecture

2.3. Image pre-processing and lesion enhancement

Each input fundus image is first resized to a fixed spatial resolution and normalized to reduce illumination variations commonly observed across screening devices. Contrast enhancement is applied to improve the visibility of subtle DR manifestations such as micro-aneurysms and hemorrhages. Noise suppression and morphological refinement are subsequently employed to preserve lesion boundaries while suppressing background artefacts. This stage produces an enhanced retinal image that supports both deep feature extraction and lesion localization. The lesion enhancement process also generates a binary lesion mask, which is later transformed into a lesion prior map. Importantly, this lesion prior is not treated as an auxiliary output; instead, it plays a direct role in guiding representation learning, forming a foundational component of the proposed methodology.

2.4. Deep feature extraction and lesion prior modelling

Deep visual features are extracted using a pretrained EfficientNet-B0 backbone, which remains frozen during training to preserve generalizable retinal representations. For a given pre-processed fundus image I, the backbone produces a spatial feature map F, as defined in (1).

$$F = f_\theta(I) \in \mathbb{R}^{H_f \times W_f \times C} \tag{1}$$

Where H_f and W_f denote the spatial dimensions of the feature map, and C represents the channel depth. This feature map retains both semantic and spatial information relevant to DR pathology. In parallel, the enhanced lesion mask is resized to match the spatial resolution of F, forming a lesion prior map $P \in \mathbb{R}^{H_f \times W_f \times 1}$. The coexistence of F and P enables the framework to reason jointly about learned retinal semantics and explicitly localized lesion regions, which is essential for pathology-aware modelling.

2.5. Lesion-prior-guided token squeezer

To transform spatial feature maps into a compact yet pathology-focused representation, the framework employs LPG-TS, as illustrated in Figure 2. Unlike conventional pooling or patch-based tokenization, LPG-TS generates lesion-centric visual tokens by adaptively fusing backbone saliency and lesion prior information. The fusion process is governed by set of learnable gates and is formulated as in (2).

$$S = \sigma(g) \odot Conv_F(F) + (1 - \sigma(g)) \odot Conv_P(P) \tag{2}$$

Where $Conv_F(\cdot)$ and $Conv_P(\cdot)$ are 1 × 1 convolutional projection of the feature map and lesion prior map, respectively; g denotes learnable gating parameters; and $\sigma(\cdot)$ represents the sigmoid activation. This formulation enables the model to dynamically balance learned visual saliency and explicit lesion evidence, depending on image quality and lesion prominence.

Spatial softmax is applied to the fused score map S to obtain normalized attention weights. These weights are then used to aggregate spatial features into a fixed number of lesion-centric tokens. This process ensures that the resulting token sequence concentrates representational capacity on clinically relevant regions rather than uniformly sampled background areas.

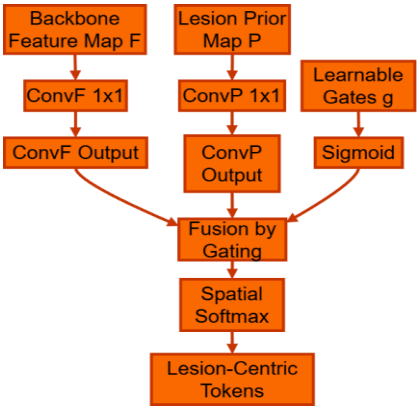


Figure 2. Schematic diagram of the LPG-TS

2.6. Handcrafted feature modelling and lesion summary token

In parallel with deep feature processing, handcrafted retinal descriptors capturing texture characteristics and lesion proxy statistics are extracted. These descriptors encode clinically interpretable information such as intensity variation, structural irregularity, and lesion component distribution. To enable interaction with transformer-based attention, the handcrafted feature vector is projected into the same latent embedding space and replicated to form a sequence of handcrafted tokens, as illustrated in Figure 3.

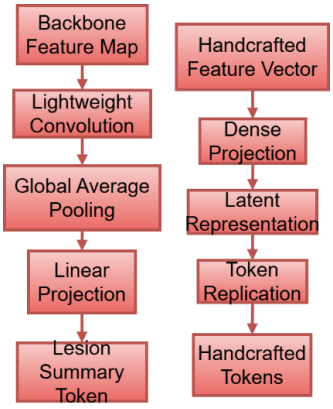


Figure 3. Lesion summary token and handcrafted token modelling

Additionally, a lesion summary token is generated from the backbone feature map using lightweight convolution, followed by global average pooling and linear projection. This token preserves global retinal context that may not be fully captured by lesion-centric tokens alone. Together, the lesion-centric tokens, lesion summary token, and handcrafted tokens form a unified multimodal token set suitable for transformer-based fusion.

2.7. Directed transformer fusion

Cross-modal integration is performed using a directed transformer fusion mechanism, as shown in Figure 4. In this design, visual tokens act as queries, while handcrafted tokens serve as keys and values. This directionality ensures that handcrafted lesion descriptors actively guide the refinement of visual representations, rather than being passively concatenated. The directed attention operation is defined as in (3).

$$Attn(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Where Q corresponds to lesion-centric visual tokens, K and V correspond to handcrafted tokens, and d_k denotes the key dimensionality. Residual connections and feed-forward refinement are applied to stabilize training and enhance representational expressiveness. This fusion strategy enables clinically meaningful cross-modal reasoning by allowing handcrafted lesion knowledge to directly influence visual attention dynamics.

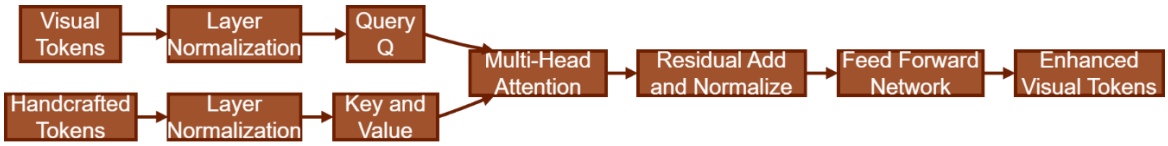


Figure 4. Directed transformer fusion mechanism

2.8. Classification and calibration assessment

The fused representation obtained after transformer fusion is passed through a dense classification head followed by a softmax layer to produce DR class probabilities. Beyond standard discrimination performance, the framework explicitly evaluates prediction confidence and calibration, as illustrated in Figure 5. Confidence scores are analyzed using expected calibration error (ECE) and Brier score to quantify the alignment between predicted probabilities and observed accuracy.

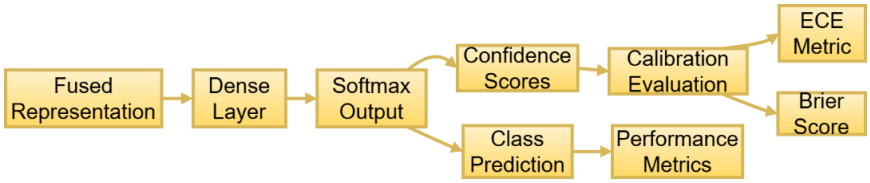


Figure 5. Classification and calibration evaluation pipeline

This calibration-aware evaluation ensures that the proposed system not only achieves high classification performance but also provides reliable confidence estimates, which are critical for real-world DR screening and referral workflows. The proposed methodology integrates lesion evidence at both the token generation level and the fusion level, ensuring that pathological information actively shapes representation learning. The LPG-TS mechanism introduces lesion-centric tokenization, while the directed transformer fusion enforces clinically guided cross-modal interaction. Together, these design choices yield a coherent, interpretable, and robust DR detection framework that aligns with practical screening requirements.

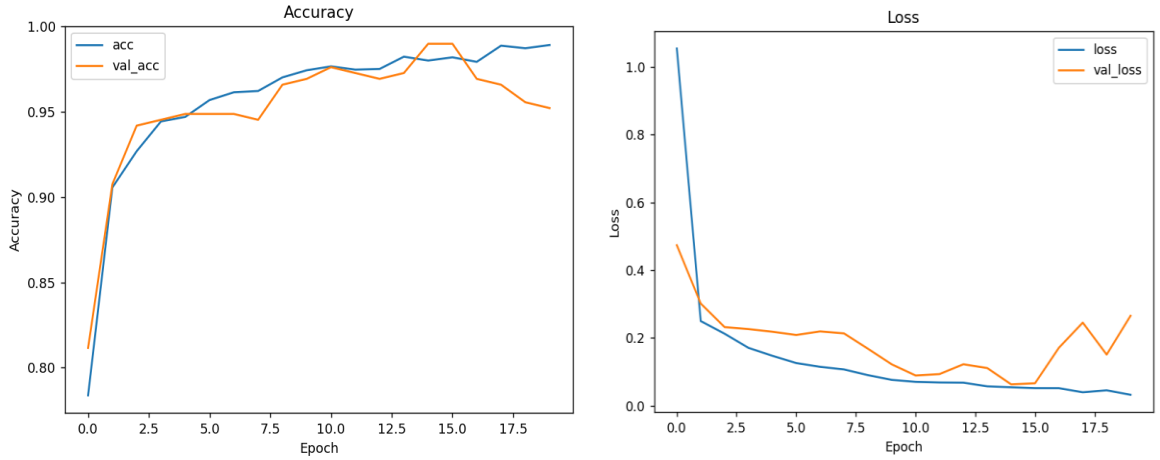
3. RESULTS AND DISCUSSION

The experimental evaluation of the proposed HandEffiTrans-DR framework is structured to analyze model behavior at three complementary levels: i) fold-wise learning dynamics and discriminative performance, ii) cross-fold robustness and calibration consistency, and iii) generalization on independent validation datasets. This multi-level analysis ensures that the reported performance is not confined to a single data split. It instead reflects stable and clinically meaningful behavior across varying data distributions.

3.1. Fold-wise training dynamics and convergence behavior

The learning behavior of the proposed model is first analyzed using the training and validation accuracy and loss curves obtained from fold-1, as shown in Figure 6. As observed in Figure 6(a), the training and validation accuracy curves exhibit a rapid increase during the initial epochs, followed by a gradual stabilization, indicating efficient feature learning and convergence. Importantly, the validation accuracy closely follows the training accuracy without significant divergence, suggesting that the model avoids overfitting despite its hybrid architecture.

Complementary evidence is provided by the loss curves in Figure 6(b), where both training and validation losses decrease monotonically and stabilize at low values. The absence of oscillatory behavior or late-epoch divergence demonstrates that the lesion-prior-guided tokenization and directed transformer fusion do not introduce training instability, even when multiple information streams are fused. This stable convergence is particularly relevant for medical image analysis, where over-parameterized models often suffer from poor generalization.



(a) (b)

Figure 6. Learning behavior of the proposed model across epochs (fold-1): (a) training and validation accuracy and (b) training and validation loss

3.2. Fold-wise discriminative performance and class-level analysis

The class-level discriminative behavior of the model for fold-1 is analyzed using the normalized confusion matrix shown in Figure 7 and the corresponding operating characteristics reported in Table 1. The confusion matrix reveals a strong diagonal dominance, with high correct classification rates for both DR and no-DR classes. Quantitatively, Table 1 reports a fold-1 accuracy of 0.9517, with a sensitivity (recall) of 0.9757 for the DR class and a specificity of 0.9134 for the no-DR class. This balance indicates that the model effectively prioritizes DR detection while maintaining reasonable specificity, which is essential for screening scenarios where false negatives are clinically costly.

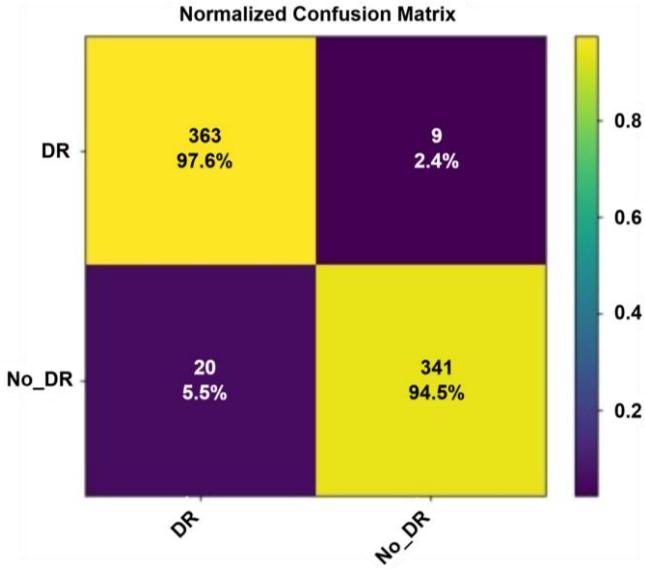


Figure 7. Normalized confusion matrix (fold-1, DR vs No-DR)

Table 1. Fold-1 operating characteristics

Metric	Formula (using DR as positive)	Value
Counts	TP = 361, FN = 9, FP = 20, TN = 211, N = 601	—
Accuracy	(TP + TN) / N	0.9517
Sensitivity / Recall (DR)	TP / (TP + FN)	0.9757
Specificity (no-DR)	TN / (TN + FP)	0.9134
Precision (DR)	TP / (TP + FP)	0.9475
F1 (DR)	2·TP / (2·TP + FP + FN)	0.9614
Precision (no-DR)	TN / (TN + FN)	0.9591
Recall (no-DR)	TN / (TN + FP)	0.9134
F1-score (no-DR)	2·TN / (2·TN + FN + FP)	0.9357
Macro-F1 (Fold-1)	Mean (F1_DR, F1_No-DR)	0.9486

The achieved performance of each class aligns well with previous lesion-aware and attention-based research in DR detection, which stated that discrimination was enhanced through highlighting lesions [16], [21]. This signifies that the developed framework successfully highlights clinically important retinal lesions. Further insight is obtained from the class-wise F1 scores. The DR class achieves an F1-score of 0.9614, while the no-DR class attains an F1-score of 0.9357, resulting in a macro-F1 of 0.9486. These results demonstrate that the model does not exhibit class bias and maintains consistent performance across both pathological and non-pathological samples. This behavior can be attributed to the lesion-centric token generation, which explicitly emphasizes pathological regions without suppressing global retinal context.

3.3. Discriminative power and probabilistic reliability

Beyond accuracy-based evaluation, the discriminative capability of the proposed framework is analyzed using receiver operating characteristic (ROC) and precision–recall (PR) curves. The ROC curve for fold-1, as shown in Figure 8, yields an AUROC of approximately 0.994, indicating excellent class separability. High AUROC values are particularly important in DR screening, as they reflect the model's ability to rank pathological images above non-pathological ones across a wide range of decision thresholds.

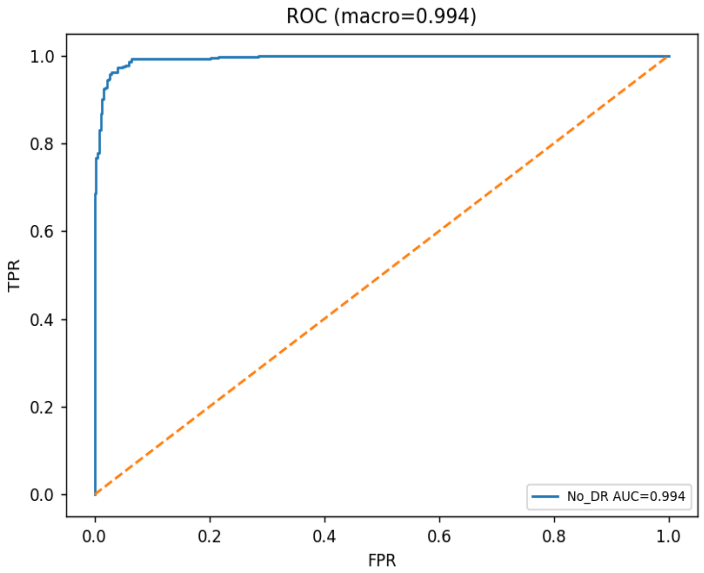


Figure 8. ROC curve with AUROC (fold-1)

Evaluation results for the HandEffiTrans-DR algorithm in terms of discrimination quality and probability calibration are shown in Figure 9. In Figure 9(a), the PR curve is depicted, whereas Figure 9(b) displays the reliability diagram. The area under the precision–recall curve (AUPRC) is quite high (AUPRC ≈ 0.994). The latter fact indicates good trade-off PR properties for the DR class. The more important point is that the reliability diagram indicates a good match between predicted probability and empirical accuracy. This observation is quantitatively supported by the discriminative and calibration metrics reported in Table 2, where the ECE and Brier score remain low (0.025 and 0.058, respectively).

Such calibration behavior is critical for clinical decision support systems, where predicted confidence often informs referral thresholds and follow-up actions. The AUROC and calibration metrics are similar to other state-of-the-art DR detector based on transformers [17], and the ECE and Brier scores are consistent with recent studies emphasizing the necessity of a well-calibrated deep learning model in healthcare decision-making systems [20], [25].

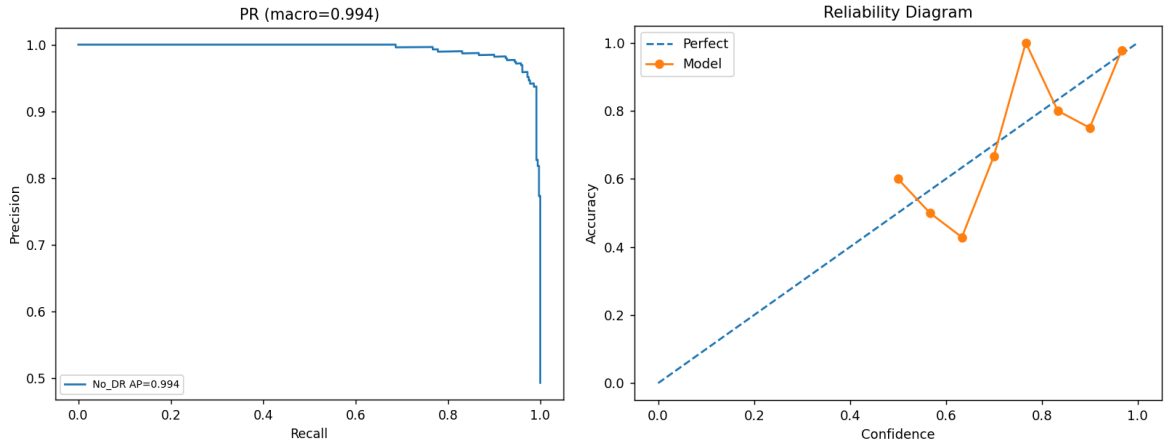


Figure 9. Performance evaluation of the proposed framework: (a) PR curve and (b) reliability (calibration)

Table 2. Fold-1 discriminative and calibration metrics

Metric	Value
AUROC	0.994
AUPRC	0.994
ECE	0.025
Brier score	0.058

3.4. Cross-fold robustness and stability analysis

To assess robustness across different data partitions, five-fold cross-validation results are summarized in Table 3. The mean accuracy across five folds is 0.9653 with a standard deviation of 0.0083, while the mean macro-F1 score is 0.9653 with a similarly low variance. These results indicate that the proposed framework maintains stable performance across folds and is not overly sensitive to specific train–test splits.

Consistently high AUROC (mean 0.9931) and AUPRC (mean 0.9937) values across folds further confirm the robustness of the discriminative capability. The relatively low standard deviation of Cohen's kappa (0.0167) suggests strong agreement beyond chance across all folds, reinforcing the reliability of the model. Importantly, calibration metrics such as ECE and Brier score remain low and stable across folds, demonstrating that probability estimates are consistently reliable and not confined to a single favorable split.

Table 3. Cross-fold performance summary (5 folds)

Metric	Mean	Std. deviation
Accuracy	0.965318	0.008345
Macro-F1	0.965303	0.008346
AUROC	0.993094	0.00209
AUPRC	0.993717	0.001382
Cohen's kappa	0.93062	0.016679
ECE	0.024945	0.007494
Brier score	0.05784	0.012628

3.5. External validation and generalization capability

To evaluate generalization beyond the primary dataset, the trained model is tested on two independent validation datasets. The performance on validation_dataset-1 is reported in Table 4, while results

on validation_dataset-2 are presented in Table 5. On validation_dataset-1, the model achieves a mean accuracy of 0.9568 and a macro-F1-score of 0.9560, with AUROC and AUPRC values exceeding 0.99. Although a slight reduction is observed compared to the primary dataset, the performance remains strong, indicating effective transferability.

Similarly, validation_dataset-2 yields a mean accuracy of 0.9627 and macro-F1 of 0.9621, with AUROC and AUPRC values of 0.9924 and 0.9928, respectively. The calibration metrics on both validation datasets remain comparable to those observed during cross-validation, suggesting that the model does not exhibit overconfident predictions when exposed to unseen data distributions. This behavior is particularly important for real-world deployment, where variations in acquisition devices and patient demographics are inevitable.

Table 4. Validation performance on validation_dataset-1

Metric	Mean	Std. deviation
Accuracy	0.95684	0.009215
Macro-F1	0.95599	0.009487
AUROC	0.99075	0.002641
AUPRC	0.9911	0.002078
Cohen's kappa	0.91567	0.018392
ECE	0.02891	0.006928
Brier Score	0.06234	0.011853

Table 5. Validation performance on validation_dataset-2

Metric	Mean	Std. deviation
Accuracy	0.96273	0.008749
Macro-F1	0.96214	0.008681
AUROC	0.99241	0.002371
AUPRC	0.99282	0.001731
Cohen's kappa	0.92458	0.017156
ECE	0.02612	0.006845
Brier Score	0.0591	0.012001

3.6. Comparative performance across datasets

A consolidated comparison of performance across the primary dataset and both validation datasets is provided in Table 6. The table highlights that the proposed framework maintains consistently high performance across all datasets, with accuracy and macro-F1 values remaining above 0.95 and discriminative metrics consistently exceeding 0.99. While a modest degradation is observed on external datasets, the drop is limited and expected, indicating strong generalization rather than dataset-specific optimization.

Table 6. Overall averaged performance comparison across datasets

Dataset	Accuracy	Macro-F1	AUROC	AUPRC	Cohen's kappa	ECE	Brier score
Primary dataset (5-fold)	0.9653	0.9653	0.9931	0.9937	0.9306	0.0249	0.0578
Validation dataset-1	0.9568	0.956	0.9908	0.9911	0.9157	0.0289	0.0623
Validation dataset-2	0.9627	0.9621	0.9924	0.9928	0.9246	0.0261	0.0591

Overall, these findings are consistent with recent CNN-transformer hybrid frameworks for retinal disease analysis [14], [17]. The proposed HandEffiTrans-DR framework further strengthens these approaches by integrating lesion-guided tokenization with directed transformer fusion, resulting in robust and reliable DR detection. The fold-wise stability, combined with strong external validation performance, supports the framework's suitability for DR screening scenarios where consistent behavior across populations and reliable confidence estimation are essential.

4. CONCLUSION

This work presented HandEffiTrans-DR, a lesion-aware hybrid framework for automated DR detection that explicitly integrates pathological priors into both representation learning and feature fusion stages. By introducing a lesion-guided tokenization strategy and a directed transformer-based fusion mechanism, the proposed approach ensures that clinically relevant lesion information actively influences the

learned visual representations. The use of stratified cross-validation and external dataset validation demonstrates robustness under class imbalance and dataset variability, while calibration analysis highlights the reliability of prediction confidence. Overall, the proposed framework advances automated DR screening by combining accuracy, robustness, and interpretability, making it suitable for large-scale clinical deployment and decision-support systems.

ACKNOWLEDGMENTS

The authors would like to acknowledge Vels Institute of Science, Technology and Advanced Studies, Chennai, India, for providing the necessary academic and research support for this study.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

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
Murali Gujjula	✓	✓	✓	✓		✓		✓	✓		✓			
Kumar Narayanan	✓	✓		✓						✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The primary retinal fundus images used in this study were obtained from the publicly accessible dataset on Kaggle at https://www.kaggle.com/c/aptos2019-blindness-detection/data. The source training set contains 3,662 labelled images: 1,805 no_DR images and 1,857 DR images after combining the mild (370), moderate (999), severe (193), and proliferative DR (295) categories. No new patient or clinical data were collected. Access to the source images requires compliance with Kaggle registration and the dataset provider's terms; therefore, the images are not redistributed with this article. Aggregate results are reported in the article. Study-generated split files and processed outputs are available from the corresponding author, [MG], upon reasonable request, subject to the original dataset terms

REFERENCES

[1] T. Y. Wong, C. M. G. Cheung, M. Larsen, S. Sharma, and R. Simó, "Diabetic retinopathy," *Nature Reviews Disease Primers*, vol. 2, no. 1, Mar. 2016, doi: 10.1038/nrdp.2016.12.

[2] M. M. Islam, H.-C. Yang, T. N. Poly, W.-S. Jian, and Y.-C. J. Li, "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis," *Computer Methods and Programs in Biomedicine*, vol. 191, Jul. 2020, doi: 10.1016/j.cmpb.2020.105320.

[3] 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, Dec. 2016, doi: 10.1001/jama.2016.17216.

[4] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, Dec. 2017, doi: 10.1001/jama.2017.18152.

[5] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, no. 1, Aug. 2018, doi: 10.1038/s41746-018-0040-6.

[6] M. Voets, K. Møllersen, and L. A. Bongo, "Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *PLOS ONE*, vol. 14, no. 6, Jun. 2019, doi: 10.1371/journal.pone.0217541.

[7] N. Tsiknakis *et al.*, "Deep learning for diabetic retinopathy detection and classification based on fundus images: a review," *Computers in Biology and Medicine*, vol. 135, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104599.

[8] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.
[9] 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, Oct. 2023, doi: 10.1371/journal.pone.0289555.
[10] G. U. Nneji, J. Cai, J. Deng, H. N. Monday, M. A. Hossin, and S. Nahar, "Identification of diabetic retinopathy using weighted fusion deep learning based on dual-channel fundus scans," *Diagnostics*, vol. 12, no. 2, Feb. 2022, doi: 10.3390/diagnostics12020540.
[11] M. T. Al-Antary and Y. Arafa, "Multi-scale attention network for diabetic retinopathy classification," *IEEE Access*, vol. 9, pp. 54190–54200, 2021, doi: 10.1109/ACCESS.2021.3070685.
[12] M. D. Alahmadi, "Texture attention network for diabetic retinopathy classification," *IEEE Access*, vol. 10, pp. 55522–55532, 2022, doi: 10.1109/ACCESS.2022.3177651.
[13] Z. Liu *et al.*, "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[14] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.
[15] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision transformer-based recognition of diabetic retinopathy grade," *Medical Physics*, vol. 48, no. 12, pp. 7850–7863, Dec. 2021, doi: 10.1002/mp.15312.
[16] 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, Mar. 2024, doi: 10.1371/journal.pone.0299265.
[17] I. U. Khan *et al.*, "A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time," *Biomedicines*, vol. 11, no. 6, May 2023, doi: 10.3390/biomedicines11061566.
[18] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, May 2021, doi: 10.3390/s21113704.
[19] R. R. -Oraá, M. H. -Tudela, M. I. López, R. Hornero, and M. García, "Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading," *Computer Methods and Programs in Biomedicine*, vol. 249, Jun. 2024, doi: 10.1016/j.cmpb.2024.108160.
[20] P. Subramaniyam and G. H. Krishnan, "Fusion learning framework for malignant and benign classification in histopathology images," in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things*, Feb. 2025, pp. 1692–1696, doi: 10.1109/IDCIOT64235.2025.10915006.
[21] B. C. Mallikarjun, K. Viswanath, B. M. Karthik, A. P. Murthy, and S. Sinha, "Retinal image analysis for detection of diabetic retinopathy - a simplified approach," *Multimedia Tools and Applications*, vol. 84, no. 8, pp. 4435–4456, Mar. 2024, doi: 10.1007/s11042-024-18821-9.
[22] 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.
[23] V. D. Raj, A. S. Sriyam, A. Meghana, H. Vutukuri, and S. Thota, "Diabetic retinopathy diagnosis using image processing and deep learning methods," in *Proceedings of the 4th International Conference on Cognitive and Intelligent Computing—Volume 1*, 2026, pp. 75–84, doi: 10.1007/978-981-95-0140-3_7.
[24] Y. Zhou *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, Oct. 2023, doi: 10.1038/s41586-023-06555-x.
[25] G. Mohandass, G. H. Krishnan, D. Selvaraj, and C. Sridhathan, "Lung cancer classification using optimized attention-based convolutional neural network with DenseNet-201 transfer learning model on CT image," *Biomedical Signal Processing and Control*, vol. 95, Sep. 2024, doi: 10.1016/j.bspc.2024.106330.

BIOGRAPHIES OF AUTHORS

Mr. Murali Gujjula is currently working as an assistant professor at Vignan's Foundation for Science, Technology and Research University and a Ph.D. scholar at Vels Institute of Science, Technology and Advanced Studies, Chennai. He obtained his M.Tech. degree in Computer Science and Engineering in affiliated of Jawaharlal Nehru Technological University, Kakinada, in 2010. He has a total of 14 years of teaching experience in the field of computer science and engineering. He has published in 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 is currently working as a professor at 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 a total of 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 international conferences. His research areas are computer networks, mobile ad hoc 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.