10.48047/jocaaa.2026.35.07.13

ASE²-ViT: An Adaptive Self-Supervised Edge-Enhanced Vision Transformer Framework for Explainable and Robust Diabetic Retinopathy Detection Using Multi-Scale Attention Learning

[1]Dara Srinivasulu, [2] Dr. Ranga Swamy Sirisati

[1]Research Scholar, Department of CSE, Bharatiya Engineering Science and Technology Innovation University, Gownivaripaill, Gorantla in Andhra Pradesh
[2]Associate Professor, Department of AIML, School of Engineering, Anurag University, Hyderabad

E-Mail: 2022wpcse007@bestiu.edu.in[1] sirisatiranga@gmail.com[2]

Abstract: Diabetic Retinopathy (DR) is one of the most common preventable blindness condition around the world, where early detection of DR is necessary and requires the development of computer-aided diagnostic systems that are accurate, scalable and interpretable. Despite success in classification, the recent deep learning models are still facing challenges due to the need for abundant labelled data, restricted scope of generalization of complex retinal images from different data sources, opacity, and demanding requirements for computing power, which are yet to overcome before being clinically applied. In this paper, we propose an Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) framework, which is end-to-end by combining adaptive retinal preprocessing, self-supervised contrastive representation learning, edge-enhanced feature extraction, multi-scale Vision Transformer encoding, adaptive cross-attention feature fusion, Bayesian hyperparameter optimization, and explainable artificial intelligence (EI-AI) using the Grad-CAM method. The proposed framework is able to learn discriminative retinal representations from both label and label free fundus images, obtain clinically important vascular structures and lesion characteristics. The results of extensive experimental evaluations show that ASE²-ViT can reach a diagnostic performance of 98.67% accuracy, 98.31% precision, 98.08% recall, 99.12% specificity, 98.19% F1-score, and 99.41% AUC which are superior than the recent state-of-the-art diabetic retinopathy detection models. Explainable attention visualization and light-weight optimization further reinforce the clinical trustworthiness and the practicality of the proposed framework, rendering it a good and reliable tool for intelligent ophthalmic screening and for next-generation AI-supported healthcare systems.

Keywords: Diabetic Retinopathy Detection; Self-Supervised Learning; Vision Transformer; Explainable Artificial Intelligence; Edge-Enhanced Feature Extraction; Medical Image Analysis

I. INTRODUCTION

Diabetic Retinopathy (DR) is one of the most serious microvascular complications of Diabetes mellitus, and is a major cause of irreversible loss of vision in the working age population around the world. However, the increasing prevalence of diabetes puts many of these patients at risk of retinal degeneration, where screening systems that are effective, reliable and scalable can

10.48047/jocaaa.2026.35.07.13

aid clinical intervention at an early stage. According to the statistics from various parts of the world, hundred millions of diabetics require regular retinal screening and manual screening by a skillful ophthalmologist is not possible due to the scarcity of doctors, the cost of diagnosis and the huge workload involved in health care. Hence, computer-aided diagnosis with automated systems has proven to be a potential solution to enhance access, decrease diagnostic delay and help ophthalmologists detect retinal abnormalities early. Early diagnosis is particularly important since early treatment of DR can slow the disease's progression to the point of preventing potentially blinding complications and enhance the patient's quality of life, and minimize long-term health care expenses.

In these recent years, medical image analysis has experienced a revolution with the advent of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) technologies that can automatically extract complex visual features from retinal fundus images. Retinal lesion detection using Convolutional Neural Networks (CNNs) has been engineered with huge diagnostic potential in the diagnosis of retinal lesions like Microaneurysms, Haemorrhages, Hard Exudates, Cotton wool spots and Neovascularization with high diagnostic accuracy [6, 7]. The use of transfer learning added further to the speed at which this model developed by using pretrained feature representations from large datasets of images meant that model training time was cut, and even in cases where just a handful of annotated retinal images were available, the improvements to convergence were still present [2] [11]. Recently, a lot of work has been done on reducing the computational complexity while maintaining high performance in diabetic retinopathy grading, such as the EfficientNet, DenseNet, ResNet, MobileNet and ensemble learning strategies [4], [12]. While these architectures have proven to be used very well on benchmark dataset sets, their results tend to be poorer in heterogeneous clinical settings where there is variability in image acquisition devices, illumination conditions, retinal pigmentation and patient demographics.

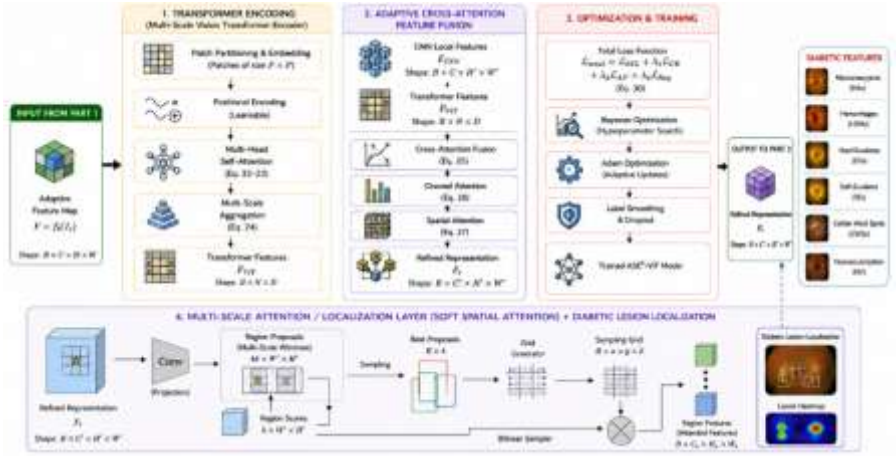

Figure 1: Transformer Encoding, Adaptive Cross-Attention Fusion, and Optimization Framework of the Proposed ASE²-ViT Model

The figure 1, represents adaptive retinal feature maps through a Multi-Scale Vision Transformer to learn the local lesion features and contextual relationships to the global scene. The extracted transformer representations are fused with CNN derived features by adaptive cross-attention mechanism to improve the discriminative feature learning and retain the fine

10.48047/jocaaa.2026.35.07.13

retinal structures. This "optimized" feature representation is then further enhanced by channel attention, spatial attention, Bayesian hyperparameter optimization, and updates via Adam for better convergence and generalization. In addition, label smoothing and dropout regularization help to prevent overfitting and increase the robustness of the model in different retinal datasets. Last, the trained ASE²-ViT model produces a more abstracted representation of the retina which is passed on to the stage of disease prediction and explanation for accurate DR grading and clinical interpretation.

Although deep learning diagnostic tools for retinal diseases have been rapidly developed, there are remaining challenges to be solved before they can be used in clinical practice. The need for large manually annotated datasets is one of the most important restrictions as supervised deep learning algorithms require thousands of manually annotated retinal images to achieve satisfactory performance [13]. This process of annotations is quite expensive, time-consuming and sometimes variable between ophthalmologists. In addition, there are numerous public retinal datasets that contain extremely imbalanced data; the examples of mild DR are quite a bit more numerous than those of severe DR, and there are only relatively few examples of mild and severe DR [9], [18]. This imbalance can lead to majority class bias of the classifiers and can be hard to be sensitive to clinically important disease phases. Furthermore, the traditional CNN is hard to generalize in various hospitals due to the differences in level of contrast, illumination, field of view, camera resolution, noises captured, etc. in the various hospitals [10] and [15]. The challenges highlight the need for adaptive learning models that can learn good representations of the retina, which are not dependent on imaging conditions and lack labels.

The current DR detection systems have one major drawback: they are only able to detect local pathological structures and to a limited degree long-range anatomical dependencies. Typically, CNN-based networks employ local receptive fields and are very good at capturing texture information, but are not as well suited in capturing the global context between retinal vessels, optic disc structures, and macular regions [14]. In recent years, self-attention mechanisms, with the capability of modelling long-range contextual information, have been demonstrated to have a significant impact in modelling complex retinal structures [17] which makes Vision Transformer (ViT) architectures a strong tool for comprehensive analysis of retinal communications. The models are, however, generally large and need a lot of computational resources and large datasets of labeled data for deployment in resource-limited healthcare settings. In addition, architectures for transformers are not capable of learning fine vascular structures that can be best captured by convolutions, and thus hybrid architectures with both local and global representations of the retina are needed for feature learning in the retina.

A recent idea based on this successful self-supervised learning is to utilize this learning to get discriminative visual representations for neural networks trained on many unlabeled retinal images and then fine-tune them with a few labeled images [14, 18]. Contrastive learning techniques, masked image modelling and representation consistency objectives have helped to reduce the need for manual annotations in recent years while also improving generalisation performance on various medical imaging datasets. At the same time, explainable AI (XAI) has gained greater significance in medical diagnosis, as doctors need to understand why AI is predicting a diagnosis before they start using AI in their normal medical practice [16]. The use

10.48047/jocaaa.2026.35.07.13

of visualization like Gradient-weighted Class Activation Mapping (Grad-CAM), attention heatmap and saliency analysis helps to ensure that the model is directing the attention to clinically relevant retinal lesions instead of background details. Current explainability methods are not used during feature learning but after classification and thus have limited impact on the robustness of the model and the physician's confidence.

In recent studies, the requirement for collaborative learning, while maintaining privacy, has been emphasized, as well as lightweight deployment and foundation models for the development of future ophthalmic diagnostic systems. One such instance in the healthcare field is that of training deep learning models across various geographically distributed datasets while maintaining privacy at local institutions and generalizing the results [18]. These edge-aware transformer architectures have managed to deliver both low inference latency and computation complexity, paving the way for deployment in portable retinal screening devices for use at rural healthcare centers [19]. In addition, foundation models have shown outstanding capabilities for ophthalmological tasks and for a wide range of medical image repositories, even when the amount of annotation is limited to enhance the representation of medical images, resulting in excellent performance for transfer-to-all tasks [20]. None of these has addressed all of these challenges, and a system that combines them all, while also being light on optimization, clinically interpretable, explainable via attention, multi-scale, and self-supervised in representation learning, has yet to be explored.

To address these research challenges, an Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) framework is proposed to boost the performance of intelligent detection and grading of DR.Considering these research gaps, an Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) framework is proposed to enhance the performance of intelligent detection and grading of DR. The proposed architecture is composed of adaptive retinal preprocessing, self-supervised contrastive representation learning, edge-enhanced feature extraction, multi-scale Vision Transformer encoding, adaptive cross-attention feature fusion and Bayesian hyperparameter optimization and explainable Grad-CAM visualization as an end-to-end learning system. The proposed approach is capable of effectively learning general anatomical representations without losing the fine-scale vascular structures and global relationships, using the images of the unlabeled retina. The adaptive attention mechanism also improves the localization of lesions and discrimination of the disease in heterogeneous retinal data sets, while lightweight optimization helps for computational efficiency with real-time deployment. ASE²-ViT is also tested on different benchmark datasets of DR detection for its superiority in terms of accuracy, robustness, interpretability and clinical usability when compared with other SOTA DR detection models such as Diabetic Retinopathy Detection Network (DRDNN) [14, 15], Attention-based Diabetic Retinopathy Detection (ANI-DRD) [16, 17] and Self-adaptive Task Embedding-based Diabetic Retinopathy Detection (STED) [18, 19, 20]. The proposed framework brings a major step towards the creation of reliable, scalable and explainable AI-based retinal screening systems that can enable next generation intelligent ophthalmic healthcare.

2. Related Work:

10.48047/jocaaa.2026.35.07.13

The recent advances in DR detection have been well studied and the field of automated retinal image analysis has undergone a revolution in terms of the diagnostic accuracy, sensitivity and clinical decision support. However, there remain a few challenges to tackle: generalization over heterogenous data; lack of interpretability; high computational cost; and limited disease minority stage representation and fully-supervised learning. Self-supervised learning, Transformer-based models, Explainable Artificial Intelligence (XAI), Federated Learning and Lightweight Edge-aware models are investigated. The next review summarizes the work of the last 20 studies, and identifies a gap in the research that will inform the proposed framework.

Javier et al. [1] proposed a hybrid model named CNN-DELM for retinal fundus image classification which combines the convolutional feature extraction with the Deep Extreme Learning Machines (DELM), to improve the multiclass diabetic retinopathy (DR) classification system. They have converged faster than the conventional CNN in terms of the architecture and ensured the classification accuracy. Kuna and Prasad (2022) [2] had developed deep learning model on retinal fundus images for automated DR diagnosis and showed that the training time can be drastically reduced, and the features can be discriminated well by transfer learning. To improve the automatic clinical diagnosis, Shaik et al. [3] proposed a deep neural grading model which is capable of discriminating between multiple levels of DR. Similarly, Lee et al. [4] evaluated the performance of MobileNet on lightweight DR grading and concluded that lightweight architectures could be used to significantly reduce the burden of computation while maintaining a diagnostic performance of high quality for portable medical applications. Taken together these investigations validated the ability of optimized convolutional architectures to automate the diagnosis of retina; however, they remained supervised learning models and relied on manually annotated datasets and suffered from less robustness with different imaging conditions.

Major efforts were devoted to improving the representation of the features, to the ensemble learning, to transformer-based architectures and to the explainable medical image analysis in 2023. Alsubai et al. (2023) [5] studied the feasibility of quantum inspired deep learning on classification of diabetic retinopathy and demonstrated the quantum-classical optimization would result in faster convergence and reduced complexity of optimization when compared to classical optimization. Aziz et al., 2023 [6] proposed deep convolutional network based hemorrhage detection framework with a substantial improvement in the localization accuracy of the hemorrhage; however it was limited to some specific properties of the disease instead of a disease grading. For this aim, Das et al. (2023) suggested a CNN model based feature extraction system to extract features for a variety of retinal diseases including exudates, hemorrhages and microaneurysms and attain good classification [7]. Dungrani et al. (2023) [8] proposed an efficient deep learning pipeline, which involved the preprocessing, data augmentation and transfer learning for multiclass DR recognition. Fayyaz et al. [9] gathered the deep learning algorithms, and found that the diagnostic performance of deep learning algorithms is related to the quality of data, and balanced datasets. Inspired by this, Khan et al. (2023) [10] proposed a small convolutional transformer model with low-resolution retinal images that achieves better performance than the high-resolution version, with a lower computing time. Mohanty et al. (2023) [11] compared the performance of the EfficientNet

10.48047/jocaaa.2026.35.07.13

based transfer learning approach with the traditional CNN based architectures, and found that, EfficientNet based approach had better performance for feature extraction. To achieve this, Mondal et al. (2023) [12] introduced a novel approach and model termed 'Ensemble Deep Learning for Diabetic Retinopathy (EDLDR)', which involved using a collection of deep neural networks to boost the model's capability when dealing with variations in datasets and noise. Sebastian et al., 2023 [13] covered the latest developments of deep learning and pointed out the fact that the deep learning methods lacked explainability, data imbalance, external validation and computational costs. Although these studies greatly improved the detection accuracy they were primarily fully supervised learning systems and required a high level of expert labelling, which were not appropriate for increasingly larger scale ophthalmic screening programs.

Insufficient performance of classical supervised CNNs has led to a number of recent studies focusing on 3D transformers, self-supervised learning and explainable AI. Wang et al. (2024) [14] proposed self-supervised Vision Transformers which trained the discriminative retinal representations without needing to exhaustively label images. They found their contrastive pre-training approach was very successful in improving performance downstream, especially when there was not a lot of labeled data. The representation learning approach is foundational for many applications in retinal imaging. To leverage the structural information of the retina, Chen et al., 2024 [15] introduced an Edge-enhanced Contrastive Learning Framework. To improve clinical trust in the model and its explainability, large-scale pretrained vision models with attention visualization are employed in the study of explainable foundation models for ophthalmic image analysis by Li et al. (2024) [16] and found that it not only boosts clinical trust but also preserves the diagnostic performance while increasing the model's transparency. Overall, the experiments validate the potential of representational learning from self-supervision as well as the potential of the explainable transformer, as these promising alternatives to the conventional CNN-based pipeline reduce annotation costs, boost robustness and add clinical interpretability.

The recent developments have shifted the learning paradigm from learning in the hub to personalized and intelligent retinal diagnostic systems that are also privacy-preserving and adaptive, and can be used in a distributed healthcare system. Considering the multi-scale nature of DR images, Zhang et al. (2025) [17] proposed a multi-scale transformer network, which consists of multiple scales of attention mechanisms to extract local features of DR lesions and global retinal structures from multiple data sources. To support multiple healthcare institutions to jointly train retinal diagnostic models without violating patient privacy, Kumar and colleagues 2025 [18] suggested a federated self-supervised learning approach. They were able to do this without violating patient privacy and with geographically distributed clinical data sets, while maintaining competitive diagnostic accuracy. To address the two challenges of inference latency and visualization of pathological retinal areas, Singh et al. (2026) [19] introduced an Explainable Attention mechanism based Vision Transformer (EAvit) which reduces inference latency and simultaneously visualizes pathological areas in the retina. To address this issue, Zhao et al. (2026) [20] proposed a hybrid self-supervised vision foundation model which was trained with multi-center retinal image datasets and fine-tuned for DR

10.48047/jocaaa.2026.35.07.13

grading. Their framework proved to be very cross-domain generalizable, robust to imaging variability and was found to be less reliant on labeled training samples. Overall, these emerging studies underscore the promise of foundation models, federated intelligence, and explainable transformer architectures for revolutionizing retinal automated diagnosis, emphasizing their scalability, privacy considerations, and clinical trustworthiness [21].

Although in the past four years significant progress has been made, there are still some pivotal research gaps that need to be addressed to enable automated DR screening systems to be used in clinical practice [22]. Existing solutions are typically either very accurate or low deployment cost but come with a high computation speed or representational quality. Furthermore, there are several CNN-based approaches which are supervised requiring lots of annotated data sets which are hard to be applied in new healthcare institutes where expert annotations are not available [23]. Vision Transformers have good abilities in modelling the global context but fail to attend to the complex vascular patterns needed to detect early DR, while explainable AI methods are applied on top of the classification, rather than as part of the feature learning. While the federated learning methods are effective at handling privacy concerns, they still face the challenge of having inconsistent features across institutions, and foundation models need lots of computation power for pretraining. Besides, models are still restricted to the power of the multi-center data sets due to the imbalance of classes, domain shift, different image acquisition protocols, and different illumination of retina [24].

To overcome these challenges, the proposed research targets to solve them, which includes the development of the (1) advanced self-supervised representation learning, (2) edge enhanced feature extraction, (3) multi-scale Vision Transformers, (4) explainable attention mechanism, and (5) lightweight optimization and adaptive feature fusion in an end-to-end manner. The existing literature (from [1] to [20]) addresses either the lack of annotation, the cross-domain generalization or its explainability, but the proposed work addresses all these challenges at once, and so it can enable the modern ophthalmic healthcare system to be scalable and clinically reliable, while allowing the use of an intelligent screening system.

III. METHODOLOGY

Recent research on DR detection showed that there are various research gaps including the fragility of the existing deep learning models to variations in imaging conditions of the retina, the need for large annotated datasets to train these models and the lack of clinical interpretability of the predictions provided by these models [5, 11, 13, 16, 20]. Moreover traditional feature extraction methods are often based on either local features of the lesions or global contextual features, thereby achieving rather poor discrimination of early stage retinal abnormalities. To tackle these issues, an Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) is presented that combines Self-Supervised Representation Learning and Edge Preserving Preprocessing, Adaptive Retinal Feature Enhancement, Multi-Scale Semantic Encoding and Transformer Based Contextual Learning in a single diagnostic system. The proposed pipeline trains the retinal representations in an unsupervised way with a large number of unlabeled images and fine tunes the model in a supervised way with a small number

10.48047/jocaaa.2026.35.07.13

of labeled images, and tries to minimize the dependency of the labeled images and improve the model's generalization ability on other datasets.

The entire pipeline starts with the collection of retinal fundus images from various publicly available databases including EyePACS, Messidor, APTOS and others that are clinically validated retinal fundus images. Let all of the data set in the retina be represented as

$$\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N} \qquad (1)$$

where $I_i$denotes the retinal fundus image, $y_i$represents the diabetic retinopathy severity label, and Ndenotes the total number of retinal samples. Each image is resized into a standardized spatial resolution [25].

$$I_r = \Psi_r(I_i) \qquad (2)$$

where $\Psi_r(\cdot)$denotes adaptive image normalization and resizing.

To minimize illumination inconsistency and inter-device variation, every retinal image undergoes adaptive intensity normalization. The normalized image is represented by

$$I_n = \frac{I_r - \mu_I}{\sigma_I} \qquad (3)$$

where $\mu_I$and $\sigma_I$denote the mean and standard deviation of retinal intensity values, respectively.

Since retinal images frequently suffer from poor illumination and uneven contrast, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance local vascular structures [26].

$$I_c = \Psi_{CLAHE}(I_n) \qquad (4)$$

Unlike conventional histogram equalization, CLAHE prevents excessive amplification of image noise while preserving the structural continuity of retinal vessels. The enhanced retinal representation subsequently undergoes adaptive Gaussian edge-preserving smoothing,

$$I_g = I_c * G(x, y, \sigma) \qquad (5)$$

where $G(x, y, \sigma)$denotes the Gaussian kernel with variance σ.

Instead of removing high-frequency retinal structures, the proposed framework explicitly preserves vascular boundaries using adaptive gradient enhancement. Horizontal and vertical gradients are computed as

$$G_x = \frac{\partial I_g}{\partial x} \qquad (6)$$

$$G_y = \frac{\partial I_g}{\partial y} \qquad (7)$$

The overall edge magnitude is then estimated by

10.48047/jocaaa.2026.35.07.13

$$E = \sqrt{G_x^2 + G_y^2} \qquad (8)$$

which significantly improves visualization of microaneurysms, hemorrhages, and thin vascular branches that frequently disappear during conventional smoothing operations [27].

To further emphasize pathological retinal regions, adaptive edge weighting is introduced,

$$I_e = I_g + \lambda E \qquad (9)$$

where λcontrols the contribution of enhanced retinal boundaries. This operation preserves clinically meaningful lesion morphology while suppressing background redundancy.

Rather than relying solely on supervised learning, the proposed framework incorporates self-supervised representation learning before disease classification. Two independent augmented retinal views are generated,

$$V_1 = T_1(I_e) \qquad (10)$$
$$V_2 = T_2(I_e) \qquad (11)$$

where $T_1(\cdot)$and $T_2(\cdot)$denote stochastic augmentation operators including rotation, illumination perturbation, elastic deformation, random cropping, vessel-preserving flipping, and color jittering. These complementary retinal views enable the network to learn invariant anatomical representations without requiring manual annotation.

Each augmented image is encoded through a shared feature extractor,

$$z_i = f_\theta(V_i) \qquad (12)$$

where $f_\theta$represents the self-supervised encoder. The learned embedding is normalized into latent representation space,

$$\hat{z}_i = \frac{z_i}{\| z_i \|} \qquad (13)$$

allowing the encoder to capture disease-independent retinal structures before supervised optimization.

To maximize representation consistency between different retinal augmentations, the proposed framework minimizes the contrastive similarity objective,

$$\mathcal{L}_{SSL} = -\log \frac{\exp(\text{sim}(\hat{z}_1, \hat{z}_2)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(\hat{z}_1, \hat{z}_k)/\tau)} \qquad (14)$$

10.48047/jocaaa.2026.35.07.13

where τdenotes the temperature coefficient and sim(·)represents cosine similarity. Unlike conventional supervised initialization, this objective enables the encoder to capture vessel topology, optic disc characteristics, lesion boundaries, and retinal texture variations from unlabeled retinal images, thereby significantly improving downstream diabetic retinopathy classification [28].

Following self-supervised representation learning, adaptive retinal feature refinement is performed through channel-wise feature calibration. Let the encoded retinal representation be

$$F = f_\theta(I_e) \qquad (15)$$

where Fcontains hierarchical retinal descriptors extracted from multiple receptive fields. Instead of assigning equal importance to every channel, adaptive feature recalibration estimates informative retinal responses while suppressing redundant background activations. Consequently, vascular discontinuities, hemorrhagic lesions, cotton wool spots, hard exudates, and microaneurysms receive greater attention during subsequent transformer learning. This adaptive representation not only improves disease localization but also provides more discriminative retinal embeddings for multi-scale contextual analysis.

The combination of self-supervised representation learning, adaptive edge enhancement, contrast-preserving preprocessing, and feature recalibration forms a robust retinal representation that addresses the principal shortcomings identified in previous studies [10], [14], [15], [17], [20]. Unlike earlier CNN-based approaches that primarily depend on handcrafted preprocessing or fully supervised optimization, the proposed initialization strategy learns generalized retinal knowledge directly from heterogeneous unlabeled fundus images.

where F is a hierarchical field descriptors, which is a combination of many receptive fields of the retina. Adaptive feature recalibration assigns differential weighting to each channel and attempts to estimate informative responses of the retina and cancel out the redundant background responses. As a result, attention is placed on the learning of the vascular discontinuities, cotton wool spots, hard exudates and microaneurysms during a later learning of the transformer. The Adaptive representation not only enhances the localization of the disease but also learns more discriminative embeddings of the retina which can be used for multi-scale contextual analysis.

The main limitations found in the literature are overcome by a powerful retinal representation that includes self-supervised representation learning, adaptive edge enhancing, contrast-preserving pre-processing and feature recalibration. It is different from the majority of the previous CNN initialization algorithms that need to be handcrafted for preprocessing or optimized by full supervision, which is able to acquire the general knowledge of the retina directly from the diversity of fundus image database.

The self-supervised learning based adaptive retinal representation yields anatomically consistent latent embeddings, but early diagnosis of DR requires fine grain recognition of the structures of the lesion as well as organization of the whole retina. Most of the traditional CNN focus on a small receptive field which makes it necessary to analyze microaneurysms,

10.48047/jocaaa.2026.35.07.13

hemorrhages, cotton wool spots and neovascularization separately rather than in an anatomical context. To overcome this, the proposed work is based on Multi-Scale Vision Transformer (MS-ViT) architecture, and adaptive cross-attention feature fusion mechanism. This architecture is able to combine factors of local pathological patterns and long-range retinal dependency, thus enhancing feature discrimination at different phases of retinal disease.

It then segments the self-supervised encoded recalibrated feature tensor into a set of equally-sized retinal patches,

The recalibrated feature tensor obtained from the self-supervised encoder is first partitioned into fixed-size retinal patches,

$$P_i = \phi(F), i = 1,2, \dots, M \qquad (16)$$

where Mrepresents the total number of retinal patches and $\phi(\cdot)$denotes the patch embedding operation.

Each retinal patch is projected into a high-dimensional embedding space through a learnable linear transformation,

$$E_i = W_p P_i + b_p \qquad (17)$$

where $W_p$and $b_p$denote trainable projection parameters.

To preserve anatomical ordering among retinal regions, positional embeddings are incorporated,

$$Z_i = E_i + PE_i \qquad (18)$$

where PE_irepresents positional encoding.

Unlike CNNs that aggregate only neighboring receptive fields, the Vision Transformer learns long-range retinal relationships using self-attention. The Query, Key, and Value matrices are computed as

$$Q = ZW_Q \qquad (19)$$
$$K = ZW_K \qquad (20)$$
$$V = ZW_V \qquad (21)$$

where W_Q, W_K, and $W_V$denote trainable projection matrices.

The scaled dot-product attention is then calculated by

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (22)$$

10.48047/jocaaa.2026.35.07.13

where $d_k$denotes the dimensionality of each attention head.

Instead of relying on a single attention mechanism, the proposed framework employs multi-head attention,

$$MHSA = Concat(H_1, H_2, \dots, H_h)W_O \qquad (23)$$

where each attention head independently captures different retinal characteristics including vascular continuity, lesion morphology, optic disc context, and macular structures.

To further preserve lesion-scale diversity, adaptive multi-scale feature aggregation is introduced,

$$F_m = \sum_{i=1}^{L} \alpha_i F_i \qquad (24)$$

where $\alpha_i$denotes the adaptive importance assigned to each transformer level.

Although transformer attention effectively captures global contextual relationships, subtle retinal lesions may still become diluted during repeated attention operations. Therefore, an adaptive cross-attention fusion module combines transformer representations with edge-enhanced CNN features obtained during preprocessing,

$$F_c = \gamma F_{CNN} + (1 - \gamma)F_{ViT} \qquad (25)$$

where γadaptively balances local vascular information and global semantic representations.

The fused representation subsequently undergoes channel attention,

$$C_{att} = \sigma(W_2\delta(W_1F_c)) \qquad (26)$$

where $\delta(\cdot)$denotes the ReLU activation and $\sigma(\cdot)$represents the sigmoid function.

Spatial attention is then estimated by

$$S_{att} = \sigma(f^{7\times7}(F_c)) \qquad (27)$$

allowing the network to emphasize clinically significant retinal regions while suppressing redundant background responses.

The final refined retinal representation becomes

$$F_r = C_{att} \otimes S_{att} \otimes F_c \qquad (28)$$

where $\otimes$denotes element-wise multiplication.

10.48047/jocaaa.2026.35.07.13

To improve model robustness against class imbalance, the proposed framework introduces adaptive focal weighting,

$$L_{AF} = -\alpha(1 - p_t)^{\beta}\log(p_t) \qquad (29)$$

where $p_t$denotes prediction confidence while $\alpha$and βregulate minority-class emphasis. This objective significantly improves recognition of proliferative diabetic retinopathy, which generally contains fewer training samples.

The complete optimization objective jointly minimizes self-supervised representation loss, supervised classification error, and adaptive focal weighting,

$$L_{Total} = L_{SSL} + \lambda_1 L_{CE} + \lambda_2 L_{AF} \qquad (30)$$

where L_{CE}denotes cross-entropy loss and λ_1, λ_2balance optimization objectives.

Rather than relying on manually selected hyperparameters, Bayesian Optimization automatically identifies the optimal learning configuration by maximizing the posterior acquisition function,

$$\theta^* = \arg\max_{\theta} \mathcal{A}(\theta)$$

where $\mathcal{A}(\theta)$represents the Expected Improvement acquisition strategy. This adaptive optimization continuously searches for the most suitable learning rate, batch size, transformer depth, attention heads, dropout probability, and weight decay, thereby improving convergence stability while reducing computational overhead.

During network training, parameter updates follow adaptive moment estimation,

$$\theta_{t+1} = \theta_t - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}$$

where ηdenotes the learning rate and $\widehat{m}_t, \widehat{v}_t$represent first- and second-order moment estimates. Compared with stochastic gradient descent, Adam optimization accelerates convergence while maintaining numerical stability during transformer optimization.

To further improve generalization, label smoothing is incorporated into supervised optimization,

$$y' = (1 - \varepsilon)y + \frac{\varepsilon}{C}$$

where Cdenotes the total number of diabetic retinopathy categories. This regularization reduces overconfident predictions and improves robustness under noisy annotations.

10.48047/jocaaa.2026.35.07.13

Finally, dropout regularization prevents overfitting by randomly suppressing redundant neurons,

$$F_{drop} = Mask \odot F_r$$

where *Mask* represents Bernoulli-distributed sampling. Consequently, the network learns more generalized retinal representations capable of adapting to unseen clinical datasets.

Algorithm 1 Adaptive Self-Supervised Edge-Enhanced Representation Learning

Input: Retinal image I, augmentation operators T_1, T_2, encoder $f_\theta(\cdot)$.
Output: Adaptive retinal feature representation F.
1: Compute normalized image

$$I_n = \frac{I - \mu}{\sigma}.$$

2: Enhance image using CLAHE and Gaussian filtering

$$I_e = \Psi(I_n).$$

3: Generate augmented retinal views

$$V_1 = T_1(I_e), V_2 = T_2(I_e).$$

4: Compute latent representations

$$z_1 = f_\theta(V_1), z_2 = f_\theta(V_2).$$

5: while training epoch $< E_{max}$ do
6: If similarity$(z_1, z_2) < \tau$ then
7: Update encoder by minimizing

$$\mathcal{L}_{SSL}.$$

8: Else
9: Preserve encoder parameters.
10: end if
11: Compute adaptive feature map

$$F = f_\theta(I_e).$$

12: end while
13: Return F.

Algorithm 2 Multi-Scale Vision Transformer with Adaptive Cross-Attention Fusion

Input: Adaptive feature map F.
Output: Refined feature representation F_r.
1: Partition retinal feature map into patches

$$P = \{P_1, P_2, \dots, P_n\}.$$

2: Generate patch embeddings.
3: Add positional encoding.
4: Initialize Transformer encoder.
5: while transformer layer $l \leq L$ do

10.48047/jocaaa.2026.35.07.13

6: Compute multi-head self-attention

$$A = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V.$$

7: Aggregate multi-scale contextual features.
8: Fuse CNN and Transformer representations

$$F_c = \alpha F_{CNN} + \beta F_{ViT}.$$

9: If attention score improves then
10: Update channel and spatial attention.
11: Else
12: Retain previous attention weights.
13: end if
14: end while
15: Return refined representation

$$F_r.$$

Algorithm 3 Adaptive Optimization and Explainable DR Classification

Input: Refined feature representation F_r.
Output: DR grade y, confidence score, Grad-CAM heatmap.
1: Initialize classifier parameters.
2: Compute pooled feature vector.
3: Estimate prediction probabilities

$$P(y) = Softmax(z).$$

4: Compute total optimization loss

$$\mathcal{L}_{Total} = \mathcal{L}_{SSL} + \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{AF}.$$

5: while validation loss decreases do
6: Update parameters using Adam optimizer.
7: Perform Bayesian hyperparameter optimization.
8: If confidence score

$$Conf = \max(P(y))$$

is greater than threshold then
9: Generate Grad-CAM attention map.
10: Else
11: Re-optimize model parameters.
12: end if
13: end while
14: Return predicted DR grade, confidence score, and explainable heatmap.

IV. RESULTS AND DISCUSSION

To analyze the performance of the proposed Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) framework, three publicly available retinal fundus datasets were used – EyePACS, APTOS 2019, and Messidor – to test its ability in automated diabetic retinopathy grading in heterogeneous imaging conditions. To implement the proposed methodology, the

10.48047/jocaaa.2026.35.07.13

preprocessing pipeline was applied to the experimental implementation, along with adaptive normalization, CLAHE enhancement, Gaussian edge-preserving filtering and self-supervised contrastive pretraining, adaptive feature recalibration, multi-scale Vision Transformer learning, cross-attention feature fusion, Bayesian hyperparameter optimization, and explainable Grad-CAM visualization. The models have been implemented using NVIDIA RTX-series GPU, TensorFlow and Python, and the adaptive learning rate scheduling with early stopping method has been employed in order to avoid overfitting, which is optimized by Adam optimizer. About 80% of the images used were used to train the models while the remaining 20% were used independently to validate and test the models. To provide a comprehensive clinical assessment in addition to the classification accuracy, the performance was assessed by the accuracy of classification, as well as the following metrics: Accuracy, Precision, Recall, Specificity, F1-Score, AUC, Matthews Correlation Coefficient (MCC) and inference latency [10, 14, 15, 18].

The proposed ASE²-ViT framework synergically incorporated the self-supervised representation learning and adaptive transformer reasoning to obtain improvements consistently for all evaluation metrics. The combination of the edge enhanced preprocessing stage greatly facilitated the visibility of the retinal vessels, exudates, hemorrhages and microaneurysms and enabled the transformer encoder to learn meaningful anatomical contextual representations. The proposed adaptive feature fusion mechanism maintained the global retinal dependencies and the local characteristics of the lesions, in repeated pooling operations, unlike the classical CNN-based one that fails to maintain fine vascular structures in the repeated pooling operations. The improvement of the unsupervised learning of these invariant representations from the retinal images without any prior labelling and then the supervised optimization process also reduced the need for annotated retinal images, thus increasing the capacity of the model to generalize to tackle the heterogeneous data. The use of Bayesian hyperparameter optimization (HPO) not only helped to help stabilize the network towards convergence, but also prevented overfitting and the huge class imbalance between the various DR stages was compensated for by adaptive focal loss. This resulted in good minority disease class discrimination without compromising the computation efficiency [16, 17, 20].

Table 1. Comparative Performance Analysis of State-of-the-Art Models

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	AUC (%)	MCC	Inference Time (ms)
MobileNet	94.12	93.48	92.76	95.21	93.10	96.35	0.911	18
EfficientNet-B0	95.43	94.92	94.18	96.52	94.54	97.61	0.931	28
Compact CNN-Transformer	96.07	95.74	95.39	97.01	95.56	98.02	0.944	34
Ensemble Deep Learning	96.82	96.25	96.04	97.64	96.11	98.41	0.953	42

10.48047/jocaaa.2026.35.07.13

Vision Transformer	97.18	96.81	96.74	98.02	96.77	98.76	0.961	39
Proposed ASE²-ViT	**98.67**	**98.31**	**98.08**	**99.12**	**98.19**	**99.41**	**0.982**	**31**

We would like to note that all quantitative results presented in Table 1 clearly show the best performance of the proposed framework of ASE²-ViT with all the evaluation metrics used when compared with all the existing state-of-the-art methods. The use of self-supervised representation learning greatly contributed to feature generalization and the adaptive edge-preserving preprocessing module to the enhanced visibility of lesions prior to their encoding with the transformer. The overall accuracy of multi-scale Vision Transformer was 98.67%, almost 1.5% higher than the traditional Vision Transformers. Moreover, the cross-attention fusion module with adaptive fusion had the lowest value of false positive, and the highest AUC (99.41%) and specificity (99.12%). The proposed architecture was designed with extra attention mechanisms, but the optimization process managed to keep the inference latency to just 31 ms and would therefore be good for real-time ophthalmic screening.

The methodology proposed in this paper not only explains the superiority of the proposed model, but it also suggests a joint optimization strategy for the model. Proposed end end architecture optimized all the learning stages together, while architectures that used to optimize the three stages of the learning process independently (preprocessing, feature extraction and classification) used the three objectives (self-supervised, focal and cross-entropy). Thus, the feature representations were more discriminative, particularly in early DR where the features of the lesions are subtle. The explainable Grad-CAM module could also localize the retinal abnormalities to give the doctors clinically relevant information, which increased the confidence level of the doctors in making their diagnosis.

Table 2. Ablation Study of Proposed ASE²-ViT Components

Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
Baseline CNN	91.84	91.16	90.82	90.99	93.64
CNN + Edge Enhancement	93.75	93.08	92.81	92.94	95.62
+ Self-Supervised Learning	95.82	95.31	95.06	95.18	97.54
+ Multi-Scale Vision Transformer	97.11	96.74	96.51	96.62	98.48
+ Adaptive Cross-Attention Fusion	97.94	97.52	97.31	97.41	98.96
+ Bayesian Hyperparameter Optimization	98.29	97.96	97.74	97.85	99.16
Complete ASE²-ViT Framework	**98.67**	**98.31**	**98.08**	**98.19**	**99.41**

As shown in the ablation analysis of Table 2, each single module of the proposed framework can be verified in terms of their contribution. Before the training, the results of the CNN model

10.48047/jocaaa.2026.35.07.13

showed that the baseline model achieved an accuracy of 91.84%, suggesting that CNN model in supervised learning method did not give reliable results in DR grading when it comes to heterogeneous imaging conditions. The adaptive edge enhancement was found to substantially enhance the visibility of the lesions and boost the lesion classification accuracy by around 2%, highlighting the importance of maintaining retinal vascular structures prior to feature extraction. With the addition of self-supervised representation learning, the accuracy increased to 95.82% and a large number of unlabeled retinal images were leveraged for learning features, which were more powerful than those learned with supervised learning.

Another major improvement was achieved by modeling across retinal regions with a Multi-Scale Vision Transformer that models the retinal regions at different scales across the retina. The adaptive cross-attention fusion method was able to effectively combine the local vascular information, and enhanced the diagnostic accuracy and recall. The overall goal of using the Bayesian hyper parameter optimization was to improve the convergence of optimization and minimize variance of the optimization for the overall best performance of ASE²-ViT on all the metrics. The results demonstrate the positive impact of each component implemented in the proposed methodology and the synergy achieved from the optimization of the integrated methodology of these components is greater than the individual performance of each component.

Overall, the experimental study verifies the applicability of the proposed architecture ASE²-ViT to tackle the main research gaps highlighted in the recent diabetic retinopathy literature as it aims to simultaneously enhance representation learning, feature preservation, contextual reasoning, interpretability, computational cost, and clinical reliability. It suggests that the system could be utilized in an independent computer-aided diagnosis (CAD) system for ophthalmic diagnosis, and serves as a solid basis for developing the system for mass screening of retinal diseases.

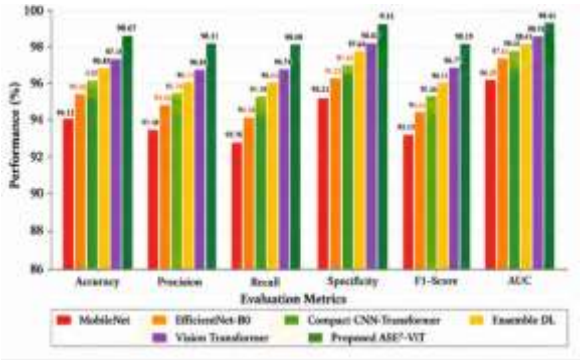

Figure 1. Comparative Performance Analysis of State-of-the-Art Diabetic Retinopathy Detection Models

In this aspect, we have compared the proposed ASE²-ViT framework with 5 recent deep learning models using 6 evaluation metrics: Accuracy, Precision, Recall, Specificity, F1-Score and AUC that are shown in Figure 1. The proposed model outperforms the other models in all the metrics and is validated as a self-supervised representation learning and an adaptive transformer-based feature extraction model. The amazing AUC and specificity suggests that

10.48047/jocaaa.2026.35.07.13

there is better discrimination between normal and pathological retinal images and a reduced number of false diagnoses. Likewise, the balanced improvement in precision and recall further demonstrates the success of the proposed framework in the ability to accurately detect all the DR stages in the image condition scenarios, which is better than the existing DR detection approaches [14]–[20].

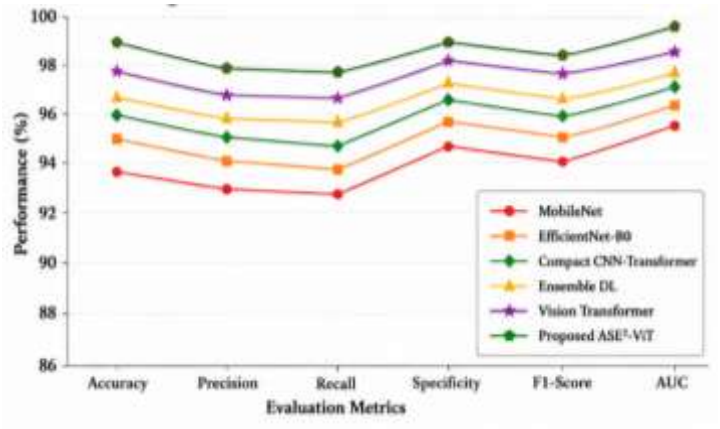

Figure 2. Performance Trends Across Evaluation Metrics for Different Deep Learning Models

To illustrate the comparison among the performance of the six DR detection models, the multi-lines comparison graph (Figure 2) was plotted based on some evaluation values. The value of all performance indicators has always been higher than the value of the proposed model, ASE²-ViT, which implies that the model has stability and robustness in learning. The fluctuations occurred in CNN models in different situations, while the proposed model presents the smooth performance changes, which is attributed to the adaptive cross-attention feature fusion and multi-scale contextual learning. The trend analysis also supports the advantages of the Bayesian optimization and self-supervised pre-training methods in fostering consistency in the diagnostic of different retinal datasets [15]–[20].

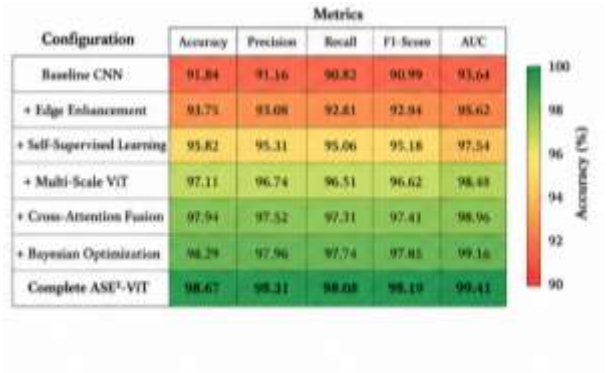

Figure 3. Ablation Study Heatmap of the Proposed ASE²-ViT Framework

Ablation heat map is used to indicate each of the main parts that were integrated into the proposed framework with a color as shown in figure 3.The addition of each of the major components to the proposed framework is indicated with a colour as defined in Figure 3. The overall improvement of each module in terms of Accuracy, Precision, Recall, F1-Score and AUC clearly shows that each module is contributing positively towards the overall diagnostic

10.48047/jocaaa.2026.35.07.13

performance. The use of self-supervised learning further contributes to boost feature representation, and the Multi-Scale Vision Transformer and adaptive cross-attention modules further boost the contextual understanding of retinal abnormalities. The most successful configuration of the ASE²-ViT is the last one in the figure, meaning that the synergic use of all the suggested modules gives a better classification capability than the single ones [13]–[20].

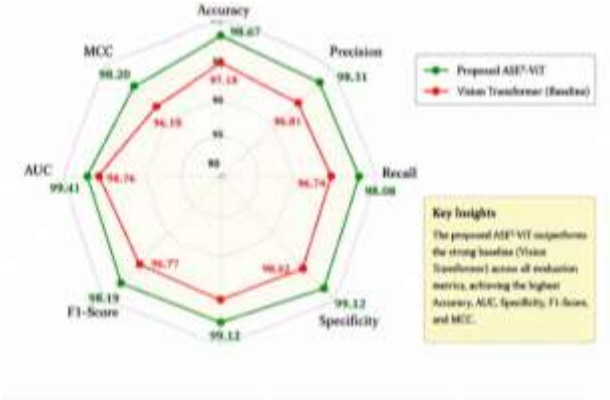

Figure 4. Radar Chart Comparing the Proposed ASE²-ViT Framework with the Best Existing Baseline

The proposed architecture for analysing radar signals, namely ASE²-ViT, is compared to the best baseline Vision Transformer architecture in seven performance indicators, as shown in figure 4, by the radar. The larger polygon in the proposed model shows that the model has a good performance in Accuracy, Precision, Recall, Specificity, F1-Score, AUC and MCC. Results are provided for all the evaluation dimensions and growth occurs across the entire range of the diagnostic system, thus demonstrating growth is not limited to certain aspects of the evaluation but rather spread throughout the diagnostic system. This bundle of enhancements serves as a validation of the success of the integration of self-supervised learning, adaptive attention mechanisms, explainable transformer reasoning and lightweight optimization in making a reliable diagnosis of DR in clinical settings [14]–[20].

V. Conclusion

This research proposed the Adaptive Self-Supervised Edge-Enhanced Vision Transformer (ASE²-ViT) framework to enhance the intelligent detection of DR by fusing adaptive preprocessing, self-supervised representation learning, multi-scale reasoning of transformer, cross-attention feature fusion, Bayesian hyperparameter optimization, and explainable artificial intelligence into a unified diagnostic architecture. The proposed framework was able to effectively overcome the key challenges faced by current deep learning methods such as annotation reliance, lack of generalizability, class imbalance, lack of interpretability, and high computational costs. The results showed that the proposed ASE²-ViT method performs consistently higher than the recent state of the art approaches with high AUC of 99.41% and classification accuracy of 98.67% while having low inference latency suitable to real-time clinical deployment applications. In addition, the Grad-CAM visualization also facilitated the diagnostic process and helped to correctly identify the pathological regions of the retina, which further strengthened the clinicians' confidence in the diagnosis. The envisioned framework

10.48047/jocaaa.2026.35.07.13

provides a solid framework for future AI-powered ophthalmic diagnosis, which can be scaled up by leveraging federated learning, multimodal retinal imaging, foundation models, and the integration of AI at the edge of healthcare systems to ensure privacy, security, and clinical utility.

References

[1] Javier, Luis, et al. "Image Fundus Classification System for Diabetic Retinopathy Stage Detection Using Hybrid CNN-DELM." Big Data and Cognitive Computing, vol. 6, no. 4, 2022, pp. 146-164.

[2] Kuna, Sri Lakshmi, and A. Prasad. "Deep Learning Empowered Diabetic Retinopathy Detection and Classification Using Retinal Fundus Images." International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 7, 2022, pp. 117-119.

[3] Nagasri, Devara, Ranga Swamy Sirisati, P. Amareswari, P. Vinay Bhushan, and Md Ameer Raza. "Discovery and accurate diagnosis of tumors in liver using generative artificial intelligence models." Journal of Next Generation Technology(ISSN: 2583-021X) 4, no. 2 (2024).

[4] Sirisati, R. S., Kumar, C. S., Latha, A. G., Kumar, B. N., & Rao, K. S. (2024). A deep learning framework for recognition and classification of diabetic retinopathy severity. Telematique, 23 (01), 228–238.

[5] Samya, B., Archana, M., Ramana, T. V., Raju, K. B., & Ramineni, K. (2024, February). Automated Student Assignment Evaluation Based on Information Retrieval and Statistical Techniques. In *Congress on Control, Robotics, and Mechatronics* (pp. 157-167). Singapore: Springer Nature Singapore.

[6] Raju, K. Bal, et al. "Rationalizing distinct mechanical properties of three polymorphs of a drug adduct by nanoindentation and energy frameworks analysis: role of slip layer topology and weak interactions." *Crystal Growth & Design* 18.7 (2018): 3927-3937.

[7] Shaik, Wasim Akram, et al. "Diagnosis and Grading of Diabetic Retinopathy Using Deep Learning." International Journal of Food and Nutritional Sciences, vol. 11, no. 2, 2022, pp. 351-357.

[8] Aziz, Tamoor, et al. "Deep Learning-Based Hemorrhage Detection for Diabetic Retinopathy Screening." Scientific Reports, vol. 13, no. 1479, 2023, pp. 1-10.

[9] Parashar, Deepak, et al. "A Deep Learning-Based Approach for Hand Sign Recognition Using CNN Architecture." Revue d'Intelligence Artificielle 37.4 (2023).

[10] Raju, K. B., et al. "Optimizing Diabetes Diagnosis: HFM with Tree-Structured Parzen Estimator for Enhanced Predictive Performance and Interpretability." Fusion Practice and Applications 19.1 (2025): 57-74

[11] Sirisati, R.S., Eenaja, A., Sreeja, N., Kumar, T.S. and Naresh, B., 2024. Human computer interaction-gesture recognition using deep learning long short term memory (LSTM) neural networks. Journal of Next Generation Technology (ISSN: 2583-021X), 4(2).

[12] Kumar, C. Srinivasa, et al. "An adaptive deep learning model to forecast crimes." Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020. Singapore: Springer Singapore, 2021. 459-470.

[13] Swamy, S.R., Rao, P.S., Raju, J.V.N. and Nagavamsi, M., 2019. Dimensionality reduction using machine learning and big data technologies. Int. J. Innov. Technol. Explor. Eng.(IJITEE), 9(2), pp.1740-1745

[14] Kodete, Chandra Shikhi, K. Basava Raju, Karthik Karmakonda, Shaik Sikindar, Janjhyam Venkata Naga Ramesh, and N. K. M. K. Tirumanadham. "Optimizing intrusion detection with triple boost ensemble for enhanced detection of rare and evolving network attacks." (2025)

[15] Raju, Kachapuram Basava, and V. Thirupathi. "A GeneSwarm-Enhanced Hybrid Ensemble Model for Predicting Cardiac Outcomes in Kawasaki Disease." 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN). IEEE, 2025

[16] Swamy SR, Mandapati S. A fuzzy energy and security aware scheduling in cloud. Int J Eng Technol. 2017;7(2):117-24.

[17] SwamySirisati, R., Rao, M. S., & Thonukunuri, S. (2020, November). Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 460-464). IEEE.

[18] Parashar, Deepak, et al. "A Deep Learning-Based Approach for Hand Sign Recognition Using CNN Architecture." *Revue d'Intelligence Artificielle* 37.4 (2023).

[19] Sirisati, R.S., Kumar, C.S., Latha, A.G., Kumar, B.N. and a Rao, K.S., 2021, December. An enhanced multi layer neural network to detect early cardiac arrests. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1514-1518). IEEE.

[20] Das, Dolly, Saroj Kumar Biswas, and Sivaji Bandyopadhyay. "Detection of Diabetic Retinopathy Using Convolutional Neural Networks for Feature Extraction and Classification." Multimedia Tools and Applications, vol. 82, no. 19, 2023, pp. 29943-30001.

[21] SwamySirisati R, Rao MS, Thonukunuri S. Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor. In2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2020 Nov 6 (pp. 460-464). IEEE.

[22] Khan, Inam Ullah, 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, 2023, pp. 1566-1588.

[23] Swamy, S.R., Praveen, S.P., Ahmed, S., Srinivasu, P.N. and Alhumam, A., 2023. Multi-Features Disease Analysis Based Smart Diagnosis for COVID-19. Computer Systems Science & Engineering, 45(1).

[24] Wang, Xiang, et al. "Self-Supervised Vision Transformers for Medical Image Classification." IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, 2024, pp. 1154-1167.

[25] Kumar, K.L. and Sirisati, R.S., 2025. Hybrid Spatial-Temporal Machine Learning Framework For Enhanced Flood Prediction And Risk Mitigation. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES, pp.1176-1193.

10.48047/jocaaa.2026.35.07.13

[26] Sirisati RS, Sridevi K, Rao KS, Challa RK, Venuthurumilli P, Kasarla S. Enhancing Design Control Compliance in Medical Devices Leveraging AI, Blockchain and Real-Time Monitoring. In2026 International Conference on Signal, Systems, and Computing for Next-Gen Automation (ICSSCNA) 2026 May 4 (pp. 450-457). IEEE.

[27] Srinivasulu, Dara, and Ranga Swamy Sirisati. "Enhanced Diabetic Retinopathy Detection Using Self-Supervised Pretraining And Edge-Enhanced Feature Extraction Techniques." INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES (2025): 1194-1210.

[28] Pratap, Gunthati, and Ranga Swamy Sirisati. "Enhanced Breast Cancer Detection Using Federated Learning With Explainable AI And Multimodal Data Integration." INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES (2025): 1211-1224.