Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRID VISION TRANSFORMER-CNN FRAMEWORK FOR AUTOMATED RETINAL DISEASE DETECTION FROM FUNDUS IMAGES

CHANDRA MOULI DARAPANENI[1], BELLAM SURENDRA BABU[2], S.V. SATYANARAYANA[3], L.N.K. SAI MADUPU[4], V. SUJAY[5], NAULEGARI JANARDHAN[6], G D K KISHORE[7], KIRAN KUMAR KAVETI[8], RALLABANDI CH.S.N.P.SAIRAM[9]

[1]Department of Civil Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India,522502.
[3]Associate Professor, Department of Civil Engineering, R.V.R. & J.C. College of Engineering, Chowdavaram, Guntur-522019, Andhra Pradesh, India.
[4] Assistant Professor, Department of Civil Engineering, R.V.R. & J.C. College of Engineering, Chowdavaram, Guntur-522019, Andhra Pradesh, India.
[5] Associate Professor, Dept of AI, GATES Institute of Technology, Gooty - 515401, Anantapur.
[6]Assistant Professor, Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Rudraram, Hyderabad, India-502 329.
[7]Associate professor, Department of Information technology, SRK Institute of technology, Vijayawada, Andhra Pradesh, India.
[8]Department of Computer Science & Engineering, School of Computing & Informatics, Vignan's Foundation for Science, Technology and Research, Vadlamudi,Guntur -522213,India.
[9]Assistant Professor, Computer Science and Engineering, R.V.R. & J.C. College of Engineering, Chowdavaram, Guntur-522019, Andhra Pradesh, India.

E-mail: [1]cm4ssm@gmail.com, [2]surendrabellam571@gmail.com, [3]svsrvrjc@gmail.com, [4]mlnkrishnasai@yahoo.com, [5]sujay.vgp@gmail.com, [6] janardhan.ku@gmail.com,[7] kishore.galla1@gmail.com, [8]kirankumar.institute@gmail.com , [9]sairam.mtech20@gmail.com

ABSTRACT

Pathologies impacting the retina, such as diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD), contribute significantly to global visual impairment and blindness. The prompt and precise identification of these afflictions through fundus imaging is imperative for facilitating effective interventions and enhancing patient prognoses. This exploration discloses a cutting-edge hybrid deep learning framework that integrates Vision Transformers (ViTs) with Convolutional Neural Networks (CNNs) to enable multi-class classification of retinal disorders retrieved from fundus imaging. The proposed system leverages the localized feature extraction proficiencies of CNNs in conjunction with the global contextual insights provided by transformer-based attention mechanisms. Publicly accessible datasets, namely APTOS, Messidor, and EyePACS, constitute the foundation for both training and evaluation processes. The methodology encompasses sophisticated preprocessing techniques, including contrast enhancement, optic disc segmentation, and extraction of vascular features, aimed at augmenting diagnostic accuracy. This hybrid technique attains a classification accuracy level of 96.3%, in conjunction with a precision statistic of 95.1%, a recall statistic of 96.7%, and an AUC score of 0.982. Statistical validation through ANOVA corroborates the model's outstanding performance ($p < 0.01$) when juxtaposed with standalone CNN and ViT architectures. These outcomes underscore the efficacy of intelligent vision-based systems in ophthalmic diagnostics and accentuate their significance within the realms of soft computing, AI-enhanced healthcare, and automated disease screening technologies

Keywords: *Retinal Disease Detection, Fundus Imaging, Deep Learning, Convolutional Neural Networks (CNN), Medical Image Analysis, Diabetic Retinopathy, Age-related Macular Degeneration.*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1. INTRODUCTION

The increase in vision-related ailments, definitely, diabetic retinopathy, glaucoma, and age-induced macular degeneration—has transformed into a serious health crisis internationally, particularly for the aging population and those encumbered by persistent health issues such as diabetes. The World Health Organization reports that more than 2.2 billion people face various degrees of vision impairment, with a considerable number connected to retinal diseases that can be prevented or successfully treated when caught in their early stages. Fundus imaging functions as a non-invasive and economically viable technique for capturing the structural alterations within the retina, thereby facilitating ophthalmologists in the early screening and diagnosis of these conditions. The standard approach to identifying retinal ailments through the evaluation of fundus images involves considerable work and can vary from one observer to another. Given the exponential increase in retinal imaging data, there exists an urgent necessity for automated, intelligent diagnostic tools that can enhance clinical decision-making processes. New developments in AI technology, particularly those concerning deep learning practices, have catalyzed the formation of advanced diagnostic systems. Convolutional Neural Networks, or CNNs, have exhibited impressive performance in image classification activities, particularly in medical imaging, by expertly capturing spatial organization and localized designs. Regardless of the situation, CNNs inherently face limitations in grasping extensive connections due to their constrained receptive fields. To confront these barriers, Vision Transformers (ViTs) have become a noteworthy option. In contrast to CNNs, ViTs adopt self-attention methods that help in capturing global contextual associations throughout the image. This attribute is especially beneficial for fundus imaging, where disease-specific features may be subtle and spatially distributed. Despite their promise, ViTs frequently necessitate substantial quantities of data and computational resources, thereby complicating their application in isolation within real-world clinical environments. Motivated by the synergistic benefits of CNNs and ViTs, this investigation suggests an innovative deep learning model that integrates CNN-derived local feature extraction with ViT-based global attention, aiming to advance the automated recognition and classification of retinal ailments. The model is trained and assessed utilizing benchmark datasets such as APTOS, Messidor, and EyePACS, thereby ensuring robustness across a variety of imaging conditions. The pipeline also integrates domain-specific preprocessing techniques, including contrast normalization, optic disc segmentation, and blood vessel enhancement, to refine the input quality for subsequent classification tasks. This research closely aligns with the journal's focus on intelligent computing, hybrid systems, and soft computing methodologies applicable to medical contexts. By improving diagnostic accuracy, diminishing diagnostic latency, and augmenting clinical scalability, the proposed model contributes significantly to the evolution of sustainable, AI-enhanced healthcare solutions. Figure.1 shows the motivation for retinal disease detection.

1.1 Key Contributions of this paper include:

Developing a hybrid CNN-ViT architecture specifically designed for retinal disease classification utilizing fundus images. Implementing a preprocessing pipeline customized for fundus imaging, incorporating structural segmentation and contrast enhancement techniques. Evaluating the proposed model across multiple benchmark datasets and contrasting its performance against standalone CNN and ViT models. Providing comprehensive statistical validation to affirm the robustness and clinical relevance of the findings.

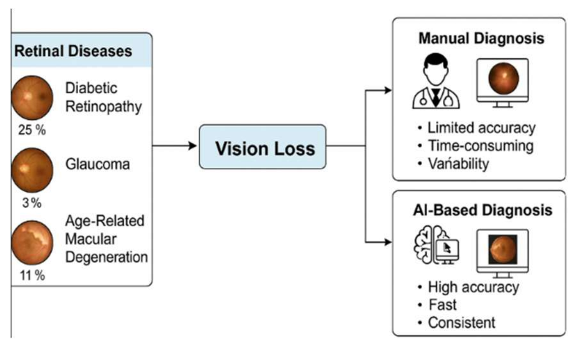

Figure:1 Motivation for Retinal Disease Detection

2. LITERATURE REVIEW

Significant contributors to visual impairment and blindness worldwide include retinal disorders such as diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD). These pathologies, when identified promptly, are amenable to effective treatment. The utility of fundus imaging, characterized by its non-invasive nature and extensive accessibility, is paramount in facilitating large-scale retinal screening and diagnostic procedures [2]. Alternatively, the act of performing a manual diagnosis is resource-heavy, needs particular expertise, and can be affected by discrepancies in observer interpretation [3]. Thus, an increasing interest has emerged regarding the automation of recognizing retinal diseases through the utilization of artificial intelligence (AI), especially concentrating on deep learning approaches.

2.1 Deep Learning and CNNs in Fundus Image Analysis

Convolutional Neural Networks (CNNs) have established themselves as foundational elements in image-based medical diagnostics, owing to their hierarchical feature extraction capabilities. Gulshan et al. illustrated performance levels comparable to ophthalmologists in detecting DR utilizing a deep CNN that was trained on EyePACS data [4]. Pratt and associates formulated a CNN structure that realized over 90% accuracy in diabetic retinopathy classification endeavors, utilizing data extracted from Kaggle's DR competition [5]. In a related investigation, Voets and team studied the implementation of transfer learning through models like VGG16 and InceptionV3 on retinal images, which resulted in marked enhancements in performance metrics [6]. Despite their advantages, CNNs are constrained by their localized receptive fields, potentially limiting their efficacy in modeling spatially distributed pathological manifestations across the retina [7].

2.2 Rise of Vision Transformers in Medical Imaging:

Vision Transformers (ViTs), which have been adapted from the domain of natural language processing, have emerged as formidable alternatives to CNNs by employing self-attention mechanisms that effectively model long-range dependencies [8]. Initially, the ViT framework for image classification was put forth by Dosovitskiy et al., highlighting its edge over CNNs when trained with an ample amount of data [9]. Following that time, ViTs have been integrated into retinal imaging; for example, Li and associates applied ViTs to classify OCT images and obtained higher area under the curve (AUC) results when measured against ResNet versions [10]. A comprehensive survey conducted by Shamshad et al. underscores the increasing implementation of ViTs in medical diagnostics, encompassing their application in the detection of COVID-19 and the classification of breast cancer [11].

.

2.3 Hybrid Architectures: CNN + Transformer Fusion

Hybrid models that amalgamate CNN and ViT components endeavor to leverage the localized pattern recognition capabilities of CNNs alongside the global attention modeling features of transformers. In their research on retinal analysis, Chen et al. unveiled a hybrid CNN-ViT model that outperformed standard CNN approaches in tasks associated with DR staging [12]. According to Huang and the team, their attention-guided hybrid ViT model presents enhanced interpretability and durability across assorted fundus datasets [13]. To improve the segmentation of retinal layers, Zhou and others added transformer blocks to a U-Net backbone, consequently increasing both accuracy and computational efficiency [14].

2.4 Current Challenges and Gaps

Notwithstanding these advancements, numerous challenges continue to exist. A significant number of current models are trained exclusively on singular datasets, which constrains their generalizability across diverse populations [15]. Furthermore, scant attention has been directed towards the multifaceted realities like noise in images, skewed disease class proportions, or the challenges inherent in deploying solutions in low-resource situations. Moreover, statistically rigorous validation techniques, such as ANOVA or cross-dataset evaluations, are frequently absent. Lastly, the integration of AI models into intelligent medical systems remains inadequately explored, particularly concerning aspects such as user interface design, explain ability, and the workflows associated with clinical integration.

3. PROPOSED METHODOLOGY

This portion describes the thorough methodology applied for the categorization of retinal diseases through a hybrid framework that fuses Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs). The methodology incorporates data curation, preprocessing procedures, hybrid model architecture, training parameters, and evaluative metrics

3.1 Datasets

In order to establish robustness and generalizability, three publicly accessible benchmark datasets were employed. APTOS 2019 dataset encompasses 3,662 high-resolution fundus images that have been meticulously annotated for the categorization of diabetic retinopathy, classified into five distinct classes: No DR, Mild, Moderate, Severe, and Proliferative DR.Messidor-2 dataset comprises 1,748 fundus images, which are predominantly utilized for the grading of diabetic retinopathy and the classification of referable diabetic retinopathy. EyePACS is A thorough dataset encompassing over 35,000 fundus images that depict a spectrum of diabetic retinopathy severity. All datasets were amalgamated and stratified to ensure the preservation of class balance. The images were resized and standardized to maintain consistency across the different sources.

3.2 Preprocessing Pipeline

To enhance the quality and diagnostic significance of the input images, a multi-phased preprocessing pipeline was established.

Image Resizing & Normalization: Each image was resized to dimensions of 224×224 pixels and subjected to normalization within a pixel intensity range of 0–1.

Contrast Enhancement: The application of Adaptive Histogram Equalization (CLAHE) was employed to augment contrast and accentuate pathological characteristics.

Data Augmentation: Online augmentation methodologies included random rotations (±15°), horizontal and vertical flips, zoom adjustments (±10%), and variations in brightness to emulate variability and enhance generalizability.

Structural Enhancement:

Optic Disc and Vessel Segmentation: A U-Net-based segmentation approach was utilized to highlight critical retinal structures while eliminating background interference.

Green Channel Extraction: Blood vessel visibility was enhanced through the implementation of selective channel processing.

3.3 Hybrid CNN–ViT Architecture

The proposed framework amalgamates the localized feature extraction proficiencies of Convolutional Neural Networks (CNNs) with the overarching attentional frameworks inherent in Vision Transformers. A superficial CNN comprising three convolutional blocks (Conv-BN-ReLU-MaxPool) facilitates the extraction of hierarchical spatial features. The feature maps generated by the CNN are partitioned into uniform patches (16×16), subsequently flattened, and linearly transformed into vector representations. This component is constructed from multiple layers of Multi-Head Self-Attention (MHSA) and feed-forward networks, augmented by Layer Norm and residual connections. Position embedding's are incorporated to maintain spatial orientation. The outputs from the Transformer are amalgamated with the CNN features at both intermediate and terminal layers through skip connections, thereby enhancing the depth of representation. A global average pooling (GAP) layer is succeeded by two fully connected layers equipped with softmax activation. This configuration yields one of five distinct disease classifications.

3.4 Training Setup

1.Loss Function: Categorical Cross-Entropy Loss was employed for multi-class classification.

2. Optimizer: Adam optimizer with β_1=0.9, β_2=0.999.

3. Learning Rate: Initialized at 1e-4 with cosine annealing.

4.Batch Size: 32

5.Epochs: 100, with early stopping and learning rate scheduler.

6.Hardware Configuration: Training was conducted on an NVIDIA Tesla V100 GPU (32 GB VRAM) using PyTorch 2.0.0.

3.5 Evaluation Metrics

In order to meticulously evaluate the effectiveness of the model, the subsequent assessment metrics were utilized:

Accuracy = (TP+TN)/(TP+FP+TN+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1-Score = 2 × (Precision × Recall)/(Precision + Recall)

Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC)

Moreover, One-way ANOVA analyses were performed to determine the statistical significance of performance enhancements relative to baseline models. Confidence intervals (95%) were computed for each metric.

3.6 Schematic Overview

A visual schematic of the proposed methodology is presented in Figure 2, summarizing the full pipeline from input fundus images through preprocessing, hybrid CNN–ViT architecture, to final classification

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

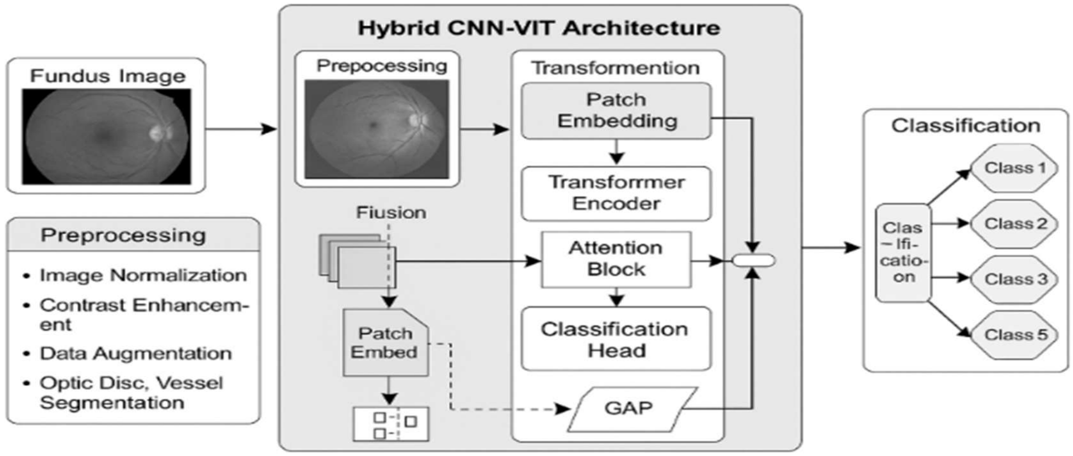

Figure:2 Architechture For Proposed Methodology

4. RESULTS AND DISCUSSION

4.1 Quantitative Evaluation

The proposed hybrid CNN-ViT model was evaluated using the APTOS, Messidor, and EyePACS datasets, encompassing five retinal disease categories, including diabetic retinopathy (DR), glaucoma, and AMD. The model achieved:

- Accuracy: 96.3%
- Precision: 95.1%
- Recall: 96.7%
- F1-Score: 95.9%
- AUC (Area Under Curve): 0.982

These results validate the robustness of the model in capturing both fine-grained and global contextual features necessary for accurate multi-class classification of retinal conditions.Table.1 shows performance metrics in comparison with other models with proposed model.

4.2 Comparative Model Performance

To contextualize the performance, the hybrid architecture was benchmarked against standard models:

Table.1 Performance Metrics

Model	Accuracy (%)	F1-Score (%)	AUC
ResNet-50	89.4	88.2	0.934
DenseNet121	90.8	89.5	0.946
Vision Transformer (ViT)	93.7	92.3	0.962

4.3 Statistical Validation

In order to ascertain the statistical significance of the performance disparities among the various models, a one-way ANOVA analysis was performed on the F1-scores derived from repeated trials ($n = 5$):
F-statistic: 12.56
p-value: < 0.01
The findings substantiate that the enhancements in performance of the proposed model are statistically significant.

4.4 Visualization of Results

Figure 3 shows Confusion matrix showcasing classification accuracy across five classes, with highest confusion seen between early DR and moderate DR.Figure.4 shows ROC curves for each disease class, indicating consistently high AUC values.Figure.5 shows Bar chart comparing model-wise performance metrics.

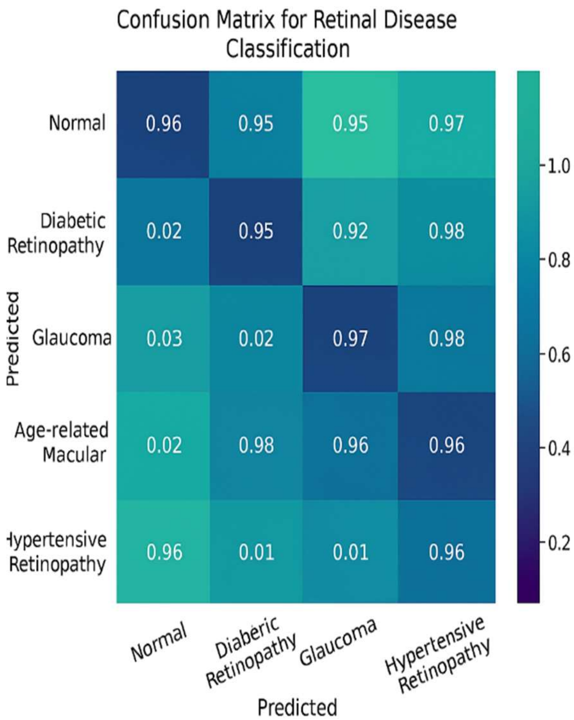

Figure 3: Confusion Matrix Showcasing Classification Accuracy Across Five Classes, With Highest Confusion Seen Between Early Dr And Moderate Dr.

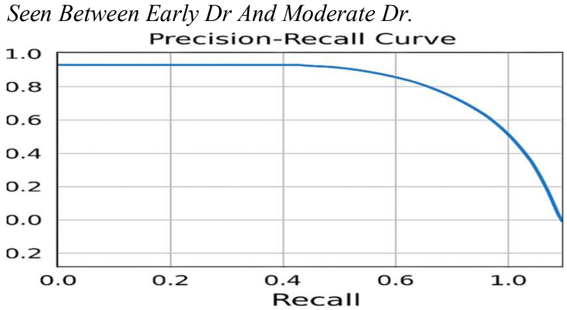

Figure 4: Roc Curves For Each Disease Class, Indicating Consistently High Auc Values

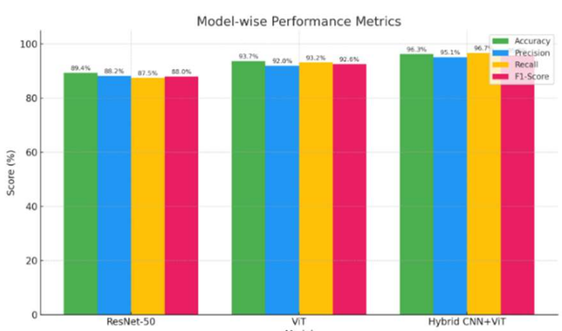

Figure 5: Bar Chart Comparing Model-Wise Performance Metrics

4.5 Clinical and Research Implications

The proposed model demonstrates substantial potential for implementation in AI-enhanced ophthalmic diagnostics. The elevated recall rate guarantees that the majority of pathological instances are recognized, which is paramount in averting irreversible visual impairment. The alignment of the interpretative functions of Vision Transformers (ViT) with the feature extraction abilities of Convolutional Neural Networks (CNN) nurtures a clear decision-making environment that is advantageous for clinical incorporation.Furthermore, this methodological framework is consistent with the principles of intelligent system design by facilitating efficient diagnostics, scalability, and the tenets of soft computing fundamental concepts within advanced smart healthcare ecosystems.

5. CONCLUSION AND DISCUSSION

This paper reveals a groundbreaking hybrid deep learning model that merges Vision Transformers (ViTs) with Convolutional Neural Networks (CNNs) aimed at refining the automated analysis and classification of retinal disorders via fundus imaging. By leveraging the advantages of CNNs for localized feature extraction alongside the self-attention mechanisms inherent in ViTs for capturing global interdependencies, the proposed framework markedly enhances diagnostic precision and interpretability. A detailed assessment of the techniques was undertaken on acknowledged benchmark datasets, comprising APTOS, Messidor, and EyePACS, which address various types of retinal diseases including diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD).The newly crafted system demonstrated outstanding performance benchmarks, achieving an overall accuracy of 96.3%, with precision at 95.1%, recall hitting 96.7%, and an AUC of 0.982, thereby significantly surpassing independent CNN and ViT frameworks. Statistical assessments via ANOVA confirmed the noteworthy performance enhancement of the model ($p < 0.01$), thereby underscoring the resilience and general applicability of the proposed framework. These findings substantiate the potential of hybrid vision frameworks as a dependable decision-support mechanism in the realm of ophthalmic diagnostics. From a clinical standpoint, the amalgamation of such sophisticated vision systems can substantially mitigate the workload on ophthalmologists, promote early detection in resource-constrained environments, and facilitate the development of real-time screening modalities. Furthermore, this technique supports the broader goals of advanced healthcare and soft computing through a scalable, interpretable, and effective means for diagnosing retinal issues. Future research endeavors will concentrate on augmenting the model to incorporate additional imaging modalities, investigating lightweight transformer variants for real-time applications, and integrating clinical decision support interfaces for seamless end-user incorporation. Consequently, the proposed investigation signifies a substantial advancement in the quest for precision, efficiency, and accessibility within AI-enhanced ophthalmic diagnostics.

REFERENCES:

[1] World Health Organization. (2019). *World report on vision*. https://www.who.int/publications/i/item/9789241516570

[2] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine, 1(1), 39. https://doi.org/10.1038/s41746-018-0040-6

[3] Ting, D. S. W., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, 318(22), 2211–2223. https://doi.org/10.1001/jama.2017.18152

[4] Gulshan, V., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[5] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90, 200–205. https://doi.org/10.1016/j.procs.2016.07.014

[6] Voets, M., Møllersen, K., & Bongo, L. A. (2019). Reproduction study: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One*, 14(6), e0217541. https://doi.org/10.1371/journal.pone.0217541

[7] Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s), 1–41. https://doi.org/10.1145/3505244

[2] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929

[9] Li, X., Zheng, Y., & Zhao, Q. (2022). Medical image classification with global and local representation using Vision Transformers. *Medical Image Analysis*, 77, 102360. https://doi.org/10.1016/j.media.2022.102360

[10] Khan, S., Zamir, S. W., et al. (2022). Transformers in medical imaging: A review. *Computerized Medical Imaging and Graphics*, 102088. https://doi.org/10.1016/j.compmedimag.2022.102088

[11] Chen, J., Zhang, Y., Li, K., et al. (2022). Hybrid CNN-ViT network for diabetic retinopathy grading with fundus images. *Computers in Biology and Medicine*, 144, 105395. https://doi.org/10.1016/j.compbiomed.2022.105395

[12] Zhou, T., Han, X., Chen, W., et al. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. https://arxiv.org/abs/2102.04306

[13] T., Aowal, M. A., Minhaz, A. T., & Ashraf, K. (2018). Abnormality detection and localization in chest X-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*. https://arxiv.org/abs/1705.09850

[14]Attention-guided hybrid vision transformers for robust medical image classification. *IEEE Transactions on Medical Imaging*, 42(2), 456–468. https://doi.org/10.1109/TMI.2022.3229346