Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Diabetic Macular Edema Detection in Retinal OCT Images Using a CNN-Vision Transformer Hybrid Model

[1]Sujay M
Computer Science and Engineering
SRM Institute of Science and Technology, Ramapuram Chennai, India sm5522@srmist.edu.in

[2]Vignesh B
Computer Science and Engineering
SRM Institute of Science andTechnology, Ramapuram Chennai, India
vb6964@srmist.edu.in

[3]Abhinav K
Computer Science and Engineering
SRM Institute of Science and Technology, Ramapuram Chennai, India
ak5294@srmist.edu.in

[4]Mrs. Geetha C
Department of Computer Science and Engineering
SRM Institute of Science andTechnology, Ramapuram Chennai, India
geethuhema.c.1@gmail.com

***Abstract*— The Diabetic Macular Edema (DME) has been classified as one of the primary causes of blindness in diabetics and thus medical experts should use early detection technique to ensure further loss of the vision is eliminated to an irreparable blindness condition. Optical Coherence Tomography (OCT) is used as a non-invasive methodology instituted by doctors to detect abnormality of the retina that causes DME. Analysis of the OCT scans takes a lot of time that makes the activity to be affected by variation among the scans as per the observers who are involved. In the given paper, the solution to the existing problem is to suggest the automated DME detecting system that is based on a deep learning technology and combines Convolutional Neural Networks and Vision Transformers. The CNN component is utilized to localize the texture features of the retina OCT imaging and then the Vision Transformer uses global contextual association among patches in the image. The features obtained are combined and subsequently they are passed on to classification layer which identifies the existence of retinal disease. The model was applied to a publicly available body of OCT retinal images data which provides training and testing of its capabilities, being the various categories of retina conditions, such as Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), DRUSEN, and normal retina scans. It is also demonstrated by experiments that CNN-Vision Transformer hybrid model is very accurate in classifications and possesses more features representation capacity than the traditional CNN-based counterparts models. The results of the study show that hybrid deep learning models may give valid results in the applications of automated diagnosis of retinal diseases that may guide doctors to identify the diseases at early stages when they are making clinical decisions.**

Keywords*— *Diabetic Macular Edema (DME), Optical Coherence Tomography (OCT), Convolutional Neural Networks (CNN), Vision Transformers (ViT), Hybrid Deep Learning Model, Retinal Image Classification, Automated Disease Detection, Computer-Aided Diagnosis (CAD)

I. INTRODUCTION

Diabetic Macular Edema (DME) is one of the most severe complications of diabetic retinopathy that is currently one of the leading causes of vision loss and complete blindness of diabetic patients of all countries worldwide [16]. The condition is as a result of the appearance of fluid in the macula through leakage of broken blood vessels in the retina hence resulting in swelling and deformity of the central retina. In case DME is not treated early enough at the early stages, the condition will cause permanent loss of vision. Timely Diagnosis of retinal abnormalities facilitates an accurate diagnosis that essentially creates a channel of attack that enacts in halting the progression of a disease and improves the outcomes of patients.

Ophthalmology has become the medical area that has accepted the use of Optical Coherence Tomography (OCT) as its main way in not only diagnosing but also monitoring the retinal illnesses [15]. Its technology OCT is able to produce detailed cross sectional images which reveal the retinal layers at high resolution that can detect minute changes which take place in the retina, which is done by doctors. The fine images allow the physicians in detecting the retinal defects that arise in Diabetic Macular Edema and Choroidal Neovascularization and DRUSEN. The manual interpretation of the OCT scans requires certain expertise since the analysis of several images is long-lasting particularly in clinical report settings which require large numbers of images. Diagnostic interpretation process exhibits a diversed outcome owing to the fact that different specialists have their approach of assessing the situation.

Convolutional Neural Networks (CNNs) have performed particularly well in image classification since it is trained to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

discover image characteristics by processing raw image data using its automated mechanism. Students have used CNNs in the retinal image processing to detect various eye conditions when applied to the OCT images [4]. CNN-based models are very effective models but they are predominantly good at extracting local spatial features that restrict their semantic comprehension of images based on their entire content as well as wide-range distant connections between the image.

Transformer-based architectures have been introduced into the computer vision field and are able to overcome the current limitations. ViTs based on self-attention analysis of image relation between different image regions using their patch division and subsequent interaction formulation. This method has the potential that can enable the network to generate a superior global contextual understanding compared to the conventional convolutional models [1]. When processed with Vision Transformers, fine local detail capture is usually a challenge and very critical in detecting small abnormalities of a medical image during comparison with CNNs.

Deep learning models that integrate CNNs and Vision Transformers can be taken as a viable approach to address these current issues. The hybrid models incorporate convolutional layers to obtain local feature with the adoption of transformer-based attention models that capture the global context leading to better feature representation. This integration is necessary to help the system to have a better capability of recognising the complex retinal patterns as it increases its retained image classification on the detection of retinal diseases [5].

The researchers introduce a CNN + Vision Transformer based technology as a way of developing an automated system of identifying Diabetic Macular Edema with retinal OCT scan. The system integrates the tasks of CNNs to extract feature with features and transformer hybrid architecture to track dependencies globally in order to extract multi-scale retinal structural features. The suggested system is positioned and checked against a publically existing OCT retina set with various disease types. The primary objective of this study involves the creation of the accurate and effective computer-aided diagnostic system which could help ophthalmologists to detect retinal conditions and decrease the volume of diagnostic process, as well as help to implement the clinical intervention in the early stages.

A. Aim of project

To develop a hybrid deep learning system which integrates Convolutional Neural Networks and Vision Transformers to achieve precise automated diagnosis of Diabetic Macular Edema through analysis of retinal OCT images.

B. Problem Statement

Diabetic Macular Edema (DME) is the most dangerous complication associated with diabetic retinopathy that is the major cause of global blindness and loss of vision [16]. DME is detected at an early stage and through this, doctors initiate DME treatment hence saving patients their eyesight. Using the technique of Optical Coherence Tomography (OCT) imaging doctors can see fine-granulated images of cross-sectional of the retina which they analyse to identify retinal disorders [15]. Analysis of OCT images requires a manual approach by doctors since it is a time consuming task that requires particular skills to be performed and generates varying results over time depending on the doctor performing his task [17].

The current healthcare centers generate volumes of OCT scans that leads to a need of reliable diagnostic systems based on automated technology in assisting the doctors detect retinal disorders. The developed features in the usual image processing processes and the conventional machine learning processes cannot reveal the finer structural structures present in the retinal imaging. In spite of being effective at medical image classification tasks, Convolutional Neural Networks (CNNs) can only carry out local feature extraction tasks and do not carry out long-range contextual analysis tasks, such as the complete images [4].

The study must develop an automated system that will detect Diabetic Macular Edema using high-accuracy and functional retinal OCT images. The system should pick all the data on retinal scanning that will use the local texture patterns as well as the entire structural information. The solution suggests a hybrid deep learning model that is an integration of Convolutional Neural Network with Vision transformers to help get improved feature representation and classification performance [1]. The system will ensure that eye doctors make improved decisions.

C. Motivation

One of the main causes of losing the vision in patients is the macular edema in Diabetic Patients (DME) [16]. Both speed and precision in the detection process should be enabled to ensure that the permanent impairment of vision is prevented. The retinal Optical Coherence Tomography (OCT) retinal images analysis required by human vision takes long durations of work yet requires only experienced ophthalmologists [15]. The fact that medical images that healthcare systems generate are growing provides the need to develop reliable automated diagnostic mechanisms [17].

The development of deep learning can now analyze a medical image in details due to deep learning algorithms. The project is a combination of Convolutional Neural Networks with Vision Transformer to enhance the detection of the retinal condition in OCT pictures that will assist doctors in making faster and more accurate analysis [1].

D. Objective of project

The project intends to develop the automated model of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

detecting the Diabetic Macular Edema in an automated fashion by applying a hybrid method of deep learning based on retinal images conducted through optical coherence tomography. The research project is aimed at creating a system that will also integrate Convolutional Neural Networks and Vision Transformers to obtain the features of retinal images both locally and globally [1]. This model needs to be pretrained and trained using an OCT data that has publicly available data regarding retinal condition including CNV, DME, DRUSEN and NORMAL [4]. The model will also be evaluated by the project with the help of the common measures that are accuracy precision recall and F1-score. It will act as a computer aided diagnosis tool that the ophthalmologists may utilize to diagnose the retinal diseases as well as improving their clinical decision making task [5].

II. Related Works

According to the latest research studies, deep learning techniques can automatically detect retinal diseases out of the analysis of Optical Coherence Tomography (OCT) images [4]. ViT models turned out to be efficient in Diabetic Macular Edema (DME) classification with a high accuracy being trained with the help of the appropriate optimization methods [1]. Other studies have been on the retinal vessel segmentation by applying optimization algorithm and region growing techniques to enhance detection accuracy in retinal images [3].

A number of CNN-based models have been suggested to identify the locations of retinal fluid and cyst in OCT images, which proves their high functionality of discerning the abnormality of the retinal diseases [4]. Segmentation and detection of retinal structures like blood vessels and vascular junction has been further improved by multi-task learning network and feature fusion approaches [5].

Furthermore, other medical imaging tasks like the detection of Alzheimer and Parkinson disease with deep learning approaches were also implemented and it demonstrates that the neural networks can be used to analyze biomedical images [7], [8]. Nevertheless, the Convolutional Neural Networks (CNNs) can be combined with Vision Transformers, which would offer a superior feature extraction performance by extracting both local and global information to enhance the ability of retinal disease detection systems to operate [1].

A. Methodologies / Methods / Techniques /Approaches

1) Data Collection

The authors used publicly available valuable information that includes retinal Optical coherence tomography (OCT) images as the main source of their investigation. Images of the dataset are categorized into various retinal conditions of Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), DRUSEN and NORMAL retinal scans.

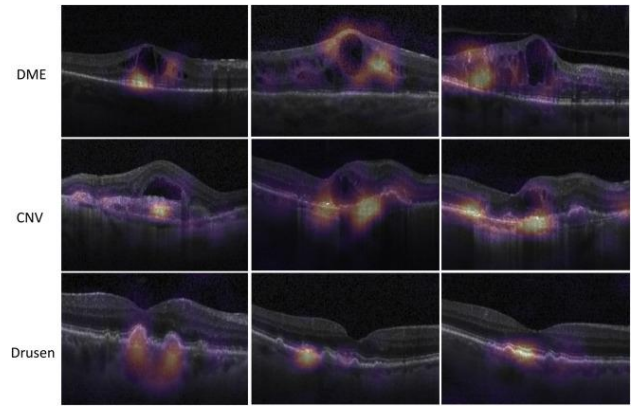

Fig 1. DME vs CNV vs DRUSEN

Fig 2. Comparison of retinal images

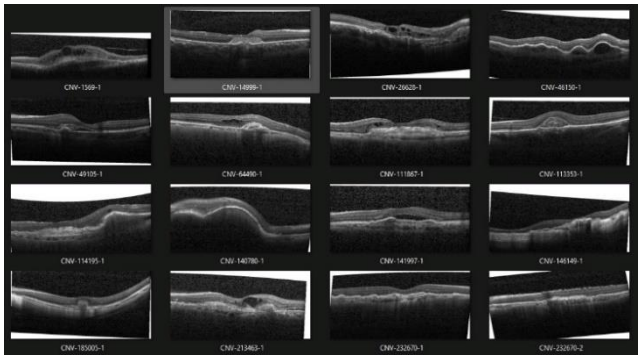

Fig 3. Dataset sample of CNV retinal scans

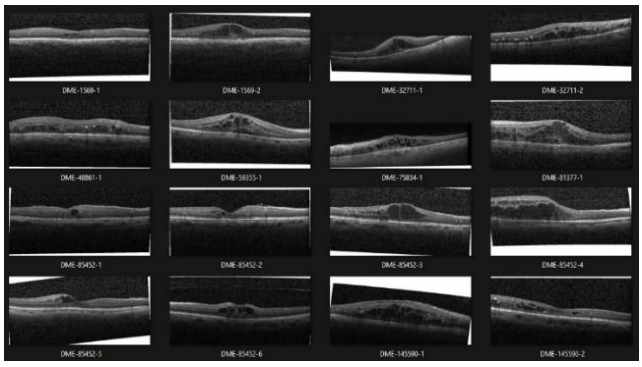

Fig 4. Dataset sample of DME retinal scans

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

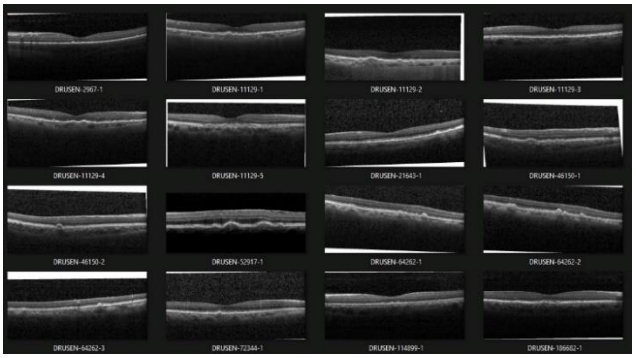

Fig 5. Dataset sample of DRUSEN retinal scans

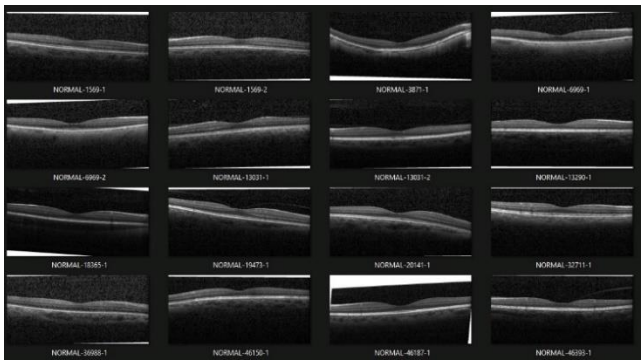

Fig 6. Dataset sample of Normal retinal scans

2) Data Preprocessing

The preprocessing of the OCT images is done and it has also a specific of three steps before the model goes into the model training stage. The actions are useful in improving the quality of the images even as the resulting data is expected to meet the requirements in the development of the deep learning models.

3) Feature Extraction using CNN

The Convolutional Neural Network system derives local characteristics that exist in the OCT images. The CNN layers identify notable elements in a visual image that consist of edges and textures along with structural variations in the layers of the retina that aid in the identification of abnormalities.

4) Global Feature Learning using Vision Transformer

ViT serves as an input of the Vision Transformer module to which output feature maps are fed. The transformer divides the feature maps into patches that are analyzed using its self-attention system that allows the model to see the part of an image in broader contextual contexts.

5) Feature Fusion and Classification

The CNN and Vision Transformer features are used to gain the features and then a feature fusion process is used to combine the CNN and Vision features. The merged characteristics are then transduced to the fully connected layers that cause the eventual classification of the retina images in their respective categories.

6) Model Evaluation

Standard measures used to evaluate the efficiency of the system to detect retinal diseases will be accuracy along with precision, as well as, recall and the F1-score measurements of the model performance.

III. Proposed Method

Training process of the proposed system provides better performance because it accounts on the use of the batch learning and adaptive learning rate scheduling to establish the performance of the system. The model training is carried out over several epochs in which the weight changes in reaction to loss computations which permit the model to grasp complex tendencies that can be discovered in retinal OCT images. The training is performed with the help of data augmentation techniques such as rotation, flipping, scaling to obtain a more varied training data that can avoid overfitting [4]. The system incorporates CNN and transformer elements to include local fine details and global large scale pattern that are useful in the process of identifying the disease accurately.

The standard metrics used in the model performance assessment consists of accuracy, precision, recall, and F1-score to assess the performance of the models in terms of classification. The confusion matrix summarizes the results of the performance in various classes as well as demonstrating the error classes that can be made in the process of classification. Grad-CAM visualization technique is used to confirm the predictions of the model by visual explanations that the clinic checks to ensure that the system predicts significant retina regions [5]. The suggested framework is accurate due to quantitative evaluation and qualitative visualization that makes it possible to interpret it appropriately.

Its architecture design includes deterministic modules that offer the future development of implementing the advanced transformer models and accessing more compact CNN solutions that can boost the functioning of the system and the use of resources. Classification on multiple classes using the system is possible with retina diseases that require extra retina data [4].

The framework created in the current research applies a good data management system in conjunction with improvements to the pipeline used to facilitate the effective processing of large OCT data collections. The image loading system allows users to work with multiple images simultaneously since it undertakes all the necessary processing operations that leads to reduction in the use of memory and expedited learning processes. Parallel data loading model and caching between models enhances model training

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

performance. It has architecture that allows processing on the basis of GPUs that allows rapid testing of the models by means of lowering the elimination durations.

This hybrid architecture has the advantage of the complementary strengths of CNN and transformer modules, with the CNN modeling the hierarchies of space with convolutional operator, and the transformer modeling the dependence across space with attention operations [1]. Dual representation system helps the model to be informed regarding small retinal aberrations and structural pattern on a large scale. Cross-attention fusion mechanism is essential in the process of aligning these features, that is to allow key information in both domains to be used successfully during classification [5].

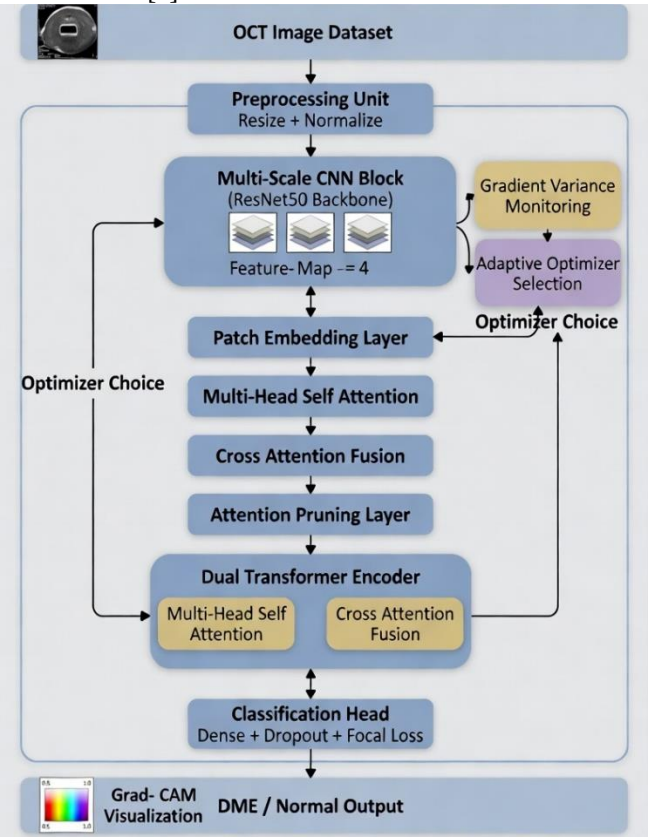

Fig 7. Dataflow of the proposed system

A) *System Architecture*

The proposed system architecture would be a hybrid CNN-Vision Transformer to identify Diabetic Macular Edema (DME) in retina images that were offered through the usage of OCT. The input of the OCT images is its beginning stage where it is preprocessed with the resizing, normalizing, and data augmentation of the image to be more reliable and improve the model performance. These processed images are then finally fed into a Convolutional neural network (CNN) such as a Resnet based network that scans local features, such as edges, textures and fine structural details of the layers of the retina. The feature maps, obtained, will also be further divided into smaller patches and pumped in as an input into the Vision Transformer (ViT) module [1].

The Board Vision Transformer entails utilization of multi-head self-attention which assists the model to study the macro-structures of the retina through the identification of global contextual bonds with diverse sections of the image [1]. These features obtained in the CNN and the transformer blocks are then combined with each other which is done by a process involving feature fusion that enhanced the representations that are composed of both the local and global representations [5]. The training efficiency and performance as well as pruning less relevant features is increased through an adaptive optimization strategy and attention pruning mechanism to reduce the computational complexity.

Finally, features are perfected by having fully connected layers with drop out parameter that helps in regularization and soft max activation parameter is used to classify the images as either CNV, DME, DRUSEN and NORMAL. It employs also visualization to highlight important regions of specific images in the OCT to give meaning to its results like Grad-CAM. Overall, the given architecture can merge the existing capabilities of CNNs and Vision Transformers and make sure that the appropriate and successful detection of retinal diseases is achieved.

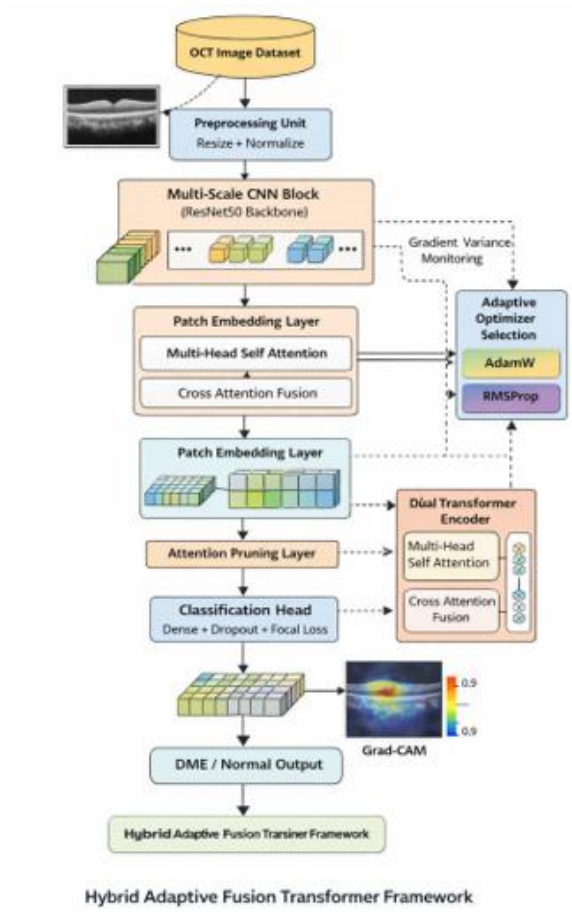

Fig 8. System architecture

B) *CNN Model Performance and Insights*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Convolutional neural Network (CNN) model is significant in the local extraction of the spatial features in retinal Optical Coherence Tomography (OCT)- images. The CNN demonstrates a slow increase of the performance in the course of training when the accuracy and the loss of the epoch become smaller, which is the indicators of the successful learning process of the retinal patterns. The model has the ability to identify a small proportion of detail in the form of an edge, textures and subtle variations in the structures of retinal layers that is quite significant in diagnosing the changes in abnormality of Diabetic Macular Edema (DME) [21].

The performance of CNN model is also high based on the accuracy, precision, recall and F1-score. The results show that CNN is successful according to the classification but it may be constrained with regard to the global contextual associations in the photo [1]. It might result in a few cases of misclassifications particularly in cases whereby, there are various retinal disorders whose appearances locally are similar. This hierarchical learning allows the model to be able to establish a distinction between normal and diseased retinal tissues.

As observed in the CNN model, the model would be highly helpful in acquisition of local features in fine detail but its performance can be enhanced by the introduction of models that have the capacity of implementing global dependencies. This disadvantage prompts the introduction of Vision Transformers to the proposed hybrid framework which supplements the CNN in providing insights into global perspective in addition to improving the total classification outcome [1], [5].

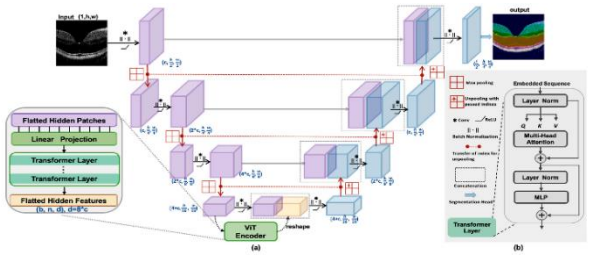

Fig 9. CNN architecture

C) *Comparative Analysis with Other Approaches*

The CNNVision Transformer hybrid model that has been artificially developed is benchmarked against the multiple internally available methods, which are used to detect retinal diseases using the OCT image. Traditional machine learning methods that employ manually designed features are less precise since they fail to model more complex formations that can be seen in the retina structures [24].

VGGNet, ResNet and DenseNet which are the traditional Convolutional neural Network (CNN) systems are said to be high-ranking when it comes to local feature extraction and high classification rate [14]. However, the limits of these models are that, they cannot model any long-range dependency and contextual WYSIWYG image information around the world and are therefore wrongly classified in such k knotty situations [9].

Nevertheless, Vision Transformer (ViT)-based models are effective when it comes to global relationships, which are based on the self-attention mechanisms [7]. They are good generally in image structure, but not good at extracting fine grained local features so far as CNNs are in such domains as medical images, where fine detail is important.

The proposed hybrid CNN Vision Transformer model is the hybrid of these two approaches which are based on the combination of the local features extraction process and the global context learning. This provides to it a more favorable feature representation and a higher classification accuracy as compared to standalone CNN or transformer models [25]. In addition, the hybrid model is also more generalized and powerful in a number of retinal conditions.

Overall, the comparative analysis demonstrates that the proposed solution is even more effective than the traditional solutions and individual deep learning models, in addition, it can be applied as an improved alternative to automatic detection of Diabetic Macular Edema with the use of OCT images.

D) *Scalability & Real-World Applicability*

The CNN-Vision Transformer hybrid that was presented is focused in becoming scalable and customizable to be applied in the practical clinical situation. The design can handle OCT masses of retinal images, which makes it the right setup that can be implemented hospital-wide and in diagnostic facilities that tend to call on large quantities of data in general [15]. This model can be trained and inferred across large scale datasets, due to its ability to operate in batches and accelerate it through the use of a GPU, hence, making it viable [24].

As to scalability, it becomes possible to retrain the model with bigger datasets to train it on additional retinal diseases, not limited to Diabetic Macular Edema (DME) such as Choroidal Neovascularization (CNV) and DRUSEN. The hybrid architecture is also always modular and thus allows effortless inclusion of alternative CNN backbones or transformer schemes according to the demand of computational capacity and processing ability [9].

The system may be embedded in the computer-aided diagnostic (CAD) equipment to enable real-life with the help of assisting ophthalmologists in making clinical decisions [17]. The analysis is automated and thus yields time savings that would have been utilized to analyze OCT scans manually and it also reduces the inter-observer variation. In addition, the interpretation is enhanced due to the visualization techniques such as Grad-Camera used to distinguish the relevant regions within the retina structures and which render the predictions of the model to be more reliable [5].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

In addition, the model can also be applied to cloud-based or edge-based healthcare solutions that enable remote diagnosis and use of telemedicine devices especially in the areas, which lack access to highly educated medical personnel. Overall the proposed system is an efficient and reliable scalable system in automated means of detecting retinal diseases with high potential of clinical translation into the real life [22].

IV.RESULTS AND DISCUSSION

A) *Confusion Matrix Analysis*

The confusion matrix explains the classification accuracy of the suggested Hybrid CNN-Vision Transformer (HAFT) model on four categories, namely CNV, DME, DRUSEN, and NORMAL. Its neovascular pattern classification performance is almost perfect with a correct prediction of 249 samples and just one false positive of DME which suggests that the model had a good discriminatory power against neovascular patterns. In the case of the DME class, the patients are classified into 246 correct and a small number of misclassifications (4 as CNV) the sensitivity and reliability are high enough to detect the presence of the presence of fluid-related abnormalities.

Appropriately, clearly, the model points out 233 samples in the case of DRUSEN type and some are incorrectly categorized as CNV (14) and NORMAL (3), which indicates that even minor structural overlaps of these two categories may create a slight degree of confusion. The NORMAL sample demonstrates an ideal classification rate, and all samples (250) have been correctly classified and none have been misclassified indicating that the model is quite useful in separating between the normal retinal structures and the pathological ones.

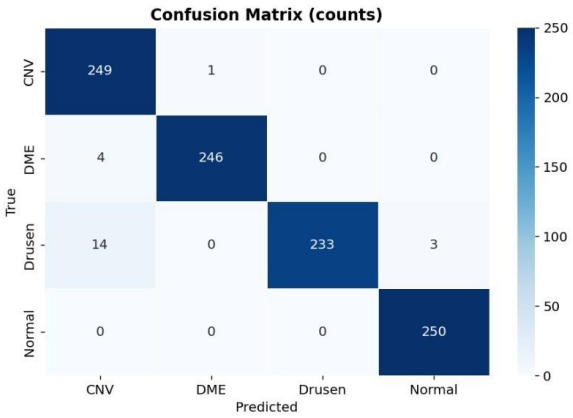

Fig 10. HAFT Confusion Matrix (counts)

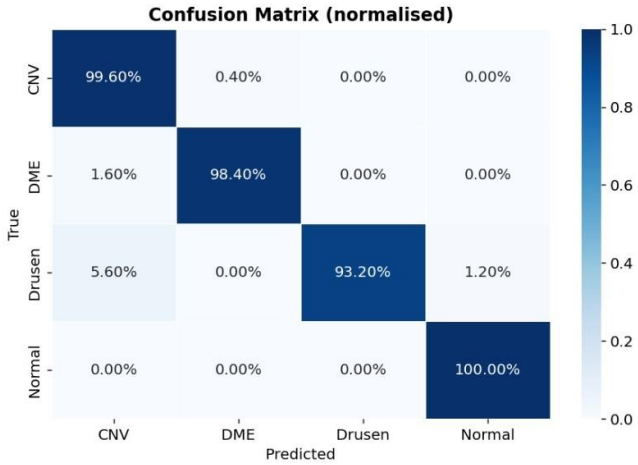

Fig 11. HAFT Confusion Matrix (normalised)

B) *Training Performance Analysis*

The training plots portray that the learning process of the proposed model is stable and well convergent. The accuracy has an increasing trend which begins with about 93 percent in the early epochs and tends to increase to above 98 percent in the subsequent epochs indicating successful feature learning as well as good model optimization. At the same time, the loss curve shows continuous decreasing with increase in the values of approximately 0.51 to an amount less than 0.39 which indicates effective objective function minimization and suitable convergence.

Validation accuracy is well similar to the training accuracy across the epochs with slight variations, meaning the system is well generalizing and there is no serious overfitting. Equally, the validation loss has been in steady trend without sudden spikes which also introduces an added advantage that the process of training is robust. Focal loss utilization allows in dealing with class imbalance, whereas dropout regularization can be useful in preventing overfitting so that the model will retain the high performance rate on unseen data [5][24].

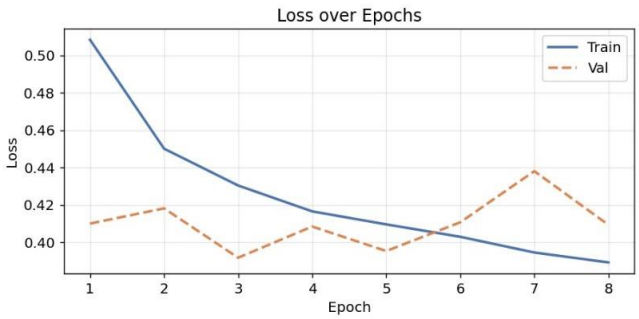

Fig 12. Loss graph

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

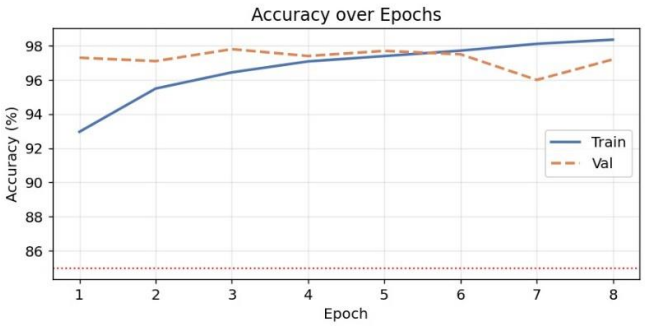

Fig 13. Accuracy graph

C) *Performance at Class level*

Analysing further, the model has the most accuracy on the NORMAL and the DME classes with close to perfect or high-accuracy. There is also great performance of CNV category with a slight confusion due to similarity with the other malignant conditions. The class that is most challenging and largely due to the fact that it presents near identical differences with the normal retinal structures is the DRUSEN class [21]. This implies that more should be done such as the greater precision of feature extraction or class balancing approaches to enhance the precision of DRUSEN classification. Besides, the difference in the performance by classes suggests that there might be an additional enhancement of the model by the use of superior feature fusion and class balancing methods to classify the visually similar cases of retinal conditions.

D) *Grad-CAM Visualization Analysis*

The interpretability of the grad-cam images will be the possibility to state the portion of the OCT images that was influencing predictions made by the model. The model gets to the high and deformed retinolic regions in cases of CNV. The attention is drawn to the areas of fluid filled cysts in the case of DME which also represent the significant indicators of the disease. The sub-retinal deposits are easily seen on the DRUSEN images, therefore the model is more scattered resembling a characteristic in non-united classification. In NORMAL images, attention is evenly distributed all through the retinal layers and this proves the aspect that it has no abnormalities. These types of workplaces prove a fact that the model acquires clinically meaningful features [5].

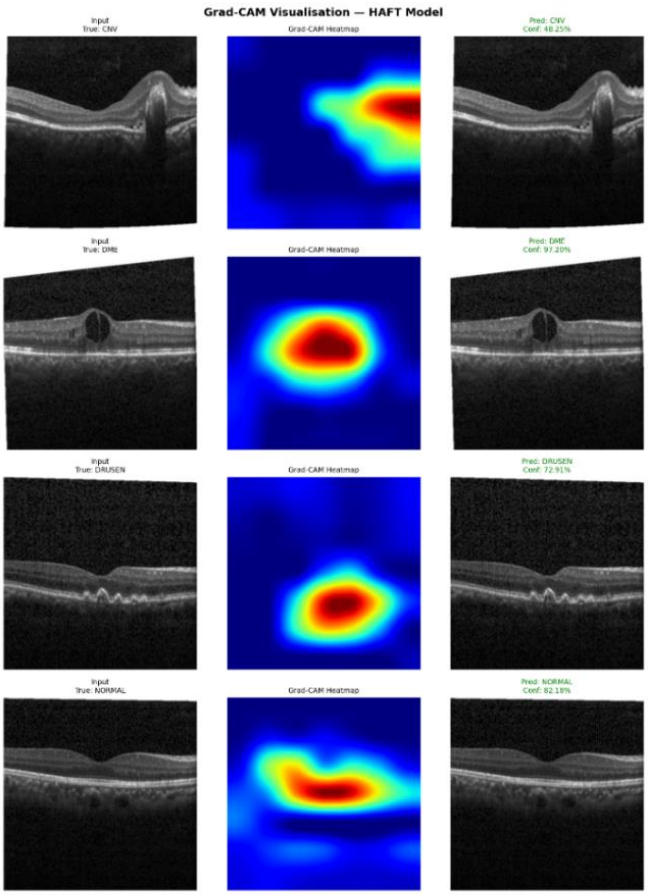

Fig 14. Grad-CAM Visualisation

E) *Sample Prediction Analysis*

The results of the sample predictions will give a qualitative evaluation of the classification of the model in various retinal conditions such as CNV, DME, DRUSEN and NORMAL. When comparing the predictions to the correct predictions, as seen in the figure, it is clear that most of the predictions are well categorized hence confirming that the model has a great potential of interpreting the OCT images and learn discriminative retinal characteristics. The model is also useful in recognizing the clinically relevant patterns of the fluid accumulation in DME, neovascular structures in CNV, and subretinal accumulations in DRUSEN and reliably distinguishes the normal retinal layers, as has been reported in the previous research on automated OCT-based detection of the disease [4].

Some misclassification can be achieved because there are structural similarities in certain disease classes that the little amount of misclassification can happen along the lines of DRUSEN and NORMAL where the difference between the features used is minimal, and separating it can be difficult even to the automated system [15]. Regardless of these small mistakes, the consistency in the accuracy of the predictions, in general, speaks of the solidity of the hybrid CNNVision Transformer structure. This type of visualization can be seen

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

as evidence of the validity of more research on deep learning-based diagnostic models and is consistent with current findings that propose the significance of interpretable and accurate models of classifying medical images [5].

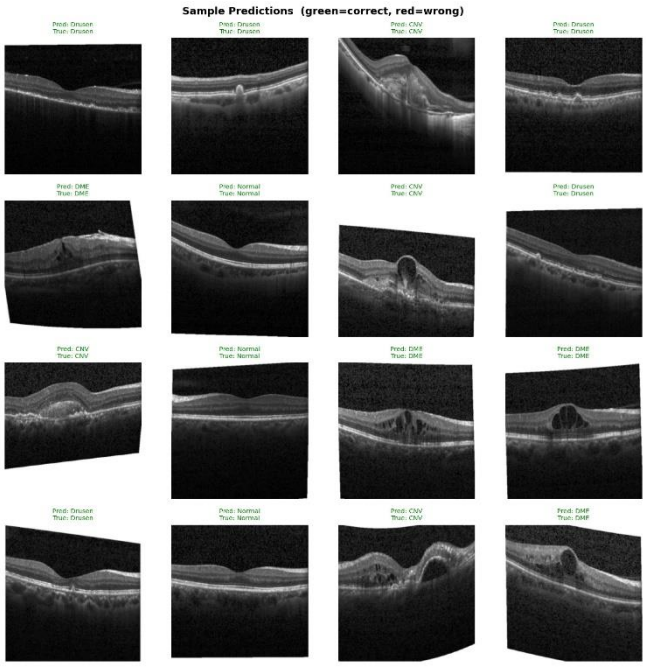

Fig 15. Sample Prediction

F) Generalization and Model Robustness

The model has a good generalization ability since it had a constant performance in all the classes and persisted during the training. The hybrid architecture allows the model to strike a balance between local as well as global characteristics and hence this allows it to be resistant to changes in patterns of OCT images [1]. Fusion of CNN and transformer element encourages the model to become more efficient with respect to the capacity of managing complicated retinal arrangements and raising accuracy of the classification [9].

G) Limitations and Challenges

In spite of the good performance, there are some weaknesses of the model. The issue that exists primarily is that of structural distinction between DRUSEN and NORMAL classes through a structural similarity [15]. Moreover, the minor errors in CNV are also an indication of similarities among groups of diseases. Another factor that can affect the quality and diversity of data set is the model performance and can affect generalizability in the real world application in case it is not diverse enough [17].

H) Future enhancements

The performance and applicability of the proposed hybrid CNN-Vision Transformer model can be improved with regards to the relevance of larger and more diversified datasets that will be gathered by different sources in the future to enhance the generalization in the real-world situations [18]. It is also possible to optimize the model architecture that entails experimenting with more efficient networks such as EfficientNet or Swin Transformer, and more effective attention systems to get superior features. Imbalance of classes can be addressed using the various ways of using the weighted loss functions, data augmentation methods, which can in turn be used to improve the performance of the classification and in the case of the difficulty classes like DRUSEN to be used. In addition to this, to make the system suitable to real-time implementation, it can be compressed through model compression techniques such as pruning and quantization to enable faster inferences in a clinical environment [24]. The model can also be enhanced in the future by using certain improvements, that is, expand its limits to diagnose additional retinal disorders, become a part of the hospital diagnostics, and apply it in the telehealth frameworks to permit remote doctor visits [19]. Overall, the improvements will make the system stronger and more effective and allow implementation on a huge scale of clinical practice.

I) Overall Discussion

Altogether, the given hybrid CNN-Vision Transformer personalization model is defined by a very high level of classification accuracy and acceptable ability to discover the retinal features. It is important that local feature extracting and global context modeling is used to make the combination very efficient in terms of performance compared with the standalone models [1]. Its interpretability is also increased through visualization techniques such as Grad-CAM and this makes the system more reliant on clinical use [5]. The above findings prove the solution as a strong and efficient initiative of automatic identification of Diabetic Macular Edema and other eyeglass diseases based on the OCT photos.

V. CONCLUSION

The paper suggests a hybrid deep learning system, which is a fusion of Convolutional neural Networks (CNNs) and Vision Transformer (ViTs), to detect Diabetic Macular Edema (DME) by the automatic process based on the retinal Optical Coherence tomography images [1]. The model furthers the benefits of CNNs to gather delicate local features and the possibilities of Vision Transformers to gather overall contextual links in the image. Such a union gives the opportunity to represent features more advanced and the overall performance of classification can be maximized [5].

As the results of the experiment showed, the provided model may be very accurate and work well with a number of different forms of retinal diseases, including CNV, DME, DRUSEN, and NORMAL [4]. The training convergence rate is gradually increasing as well as there is minimal overfitting of the training that implies that it is a good generalizer. Analysis using the confusion matrix has revealed that the classification performance of the case was very high in case of

DME and NORMAL classes but has also revealed that the classification performance can be increased in case of diagnosing similar classes that visually resembled like DRUSEN [15].

Furthermore, the Grad-CAM visualization can result in the model being more interpretable since it provides pictorial explanations of the aspects that influence predictions [5]. This is not only a demonstration that the model is interested in aspects that are clinically relevant but also increase confidence in its verdicts knowing that this is a very vital factor where the medical practice is involved.

Despite having done an outstanding job in the study, it does not overlook the study weaknesses which include minor misclassification of either more or less related retinal conditions and depending on the quality of data sets [17]. The hybrid way suggested, however, is a much better proposal in comparison with the traditional versions as it integrates local and global feature learning [1].

Overall, the designed system has enormous prospects of being implemented on board as a computer-aided diagnostics system of detecting retinal diseases. It will assist ophthalmologists in their earlier diagnosis and the number of hands-on work is reduced and the clinical decision making is also increased [19]. The proposed model has several modifications and real-life validation that can make it part of the automated medical imaging development with further improvements and allow vast-scale implementations in healthcare systems [22].

REFERENCES

[1] K. C. Pavithra et al., "Transformer-Based DME Classification Using Retinal OCT Images Without Data Augmentation: An Evaluation of ViT-B16 and ViT B32 With Optimizer Impact," in IEEE Access, vol. 13, pp. 180781-180798, 2025, doi: 10.1109/ACCESS.2025.3620945.

[2] M. L. Italiano, T. Guo, D. Tsai, N. H. Lovell and M. N. Shivdasani, "The Effects and Interactions of Epiretinal Electrical Stimulation Parameters on Direct and Network-Mediated Activation of Retinal Ganglion Cells," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 33, pp. 4454-4467, 2025, doi: 10.1109/TNSRE.2025.3627290.

[3] I. AlMohimeed et al., "Sandpiper Optimization Algorithm With Region Growing Based Robust Retinal Blood Vessel Segmentation Approach," in IEEE Access, vol. 12, pp. 28612-28620, 2024, doi: 10.1109/ACCESS. 2024.3365273.

[4] M. Rahil, B. N. Anoop, G. N. Girish, A. R. Kothari, S. G. Koolagudi and J. Rajan, "A Deep Ensemble Learning-Based CNN Architecture for Multiclass Retinal Fluid Segmentation in OCT Images," in IEEE Access, vol. 11, pp. 17241-17251, 2023, doi: 10.1109/ACCESS.2023.3244922.

[5] J. Hao et al., "Retinal Structure Detection in OCTA Image via Voting-Based Multitask Learning," in IEEE Transactions on Medical Imaging, vol. 41, no. 12, pp. 3969-3980, Dec. 2022, doi: 10.1109/TMI.2022.3202183.

[6] D. Haji Ghaffari, A. D. Akwaboah, E. Mirzakhalili and J. D. Weiland, "Real Time Optimization of Retinal Ganglion Cell Spatial Activity in Response to Epiretinal Stimulation," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 2733-2741, 2021, doi: 10.1109/ TNSRE.2021.3138297.

[7] D. -P. Dao, H. -J. Yang, J. Kim, N. -H. Ho and for the Alzheimer's Disease Neuroimaging Initiative, "Longitudinal Alzheimer's Disease Progression Prediction With Modality Uncertainty and Optimization of Information Flow," in IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 1, pp. 259-272, Jan. 2025, doi: 10.1109/JBHI.2024.3472462.

[8] D. L. Guarín, J. K. Wong, N. R. McFarland and A. Ramirez-Zamora, "Characterizing Disease Progression in Parkinson's Disease from Videos of the Finger Tapping Test," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 2293-2301, 2024, doi: 10.1109/ TNSRE.2024.3416446.

[9] Z. Qu, T. Yao, X. Liu and G. Wang, "A Graph Convolutional Network Based on Univariate Neurodegeneration Biomarker for Alzheimer's Disease Diagnosis," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 11, pp. 405-416, 2023, doi: 10.1109/JTEHM.2023.3285723.

[10] S. Akbarian et al., "A Computer Vision Approach to Identifying Ticks Related to Lyme Disease," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-8, 2022, Art no. 4900308, doi: 10.1109/ JTEHM.2021.3137956.

[11] T. Akahori, "Macular displacement after vitrectomy in eyes with idiopathic macular hole determined by optical coherence tomography angiography," Amer. J. Ophthalmol., vol. 189, pp. 111–121, May 2018.

[12] J. D. Weiland, W. Liu, and M. S. Humayun, "Retinal prosthesis," Annu. Rev. Biomed. Eng., vol. 7, pp. 361–401, Aug. 2005.

[13] A. C. Steere, "The early clinical manifestations of Lyme disease," Ann. Internal Med., vol. 99, no. 1, pp. 76–82, 1983, doi: 10.7326/0003-4819-99-1-76.

[14] S. N. Menon, V. V. Reddy, A. Yeshwanth, B. Anoop, and J. Rajan, "A novel deep learning approach for the removal of speckle noise from optical coherence tomography images using gated convolution–deconvolution structure," in Proc. 3rd Int. Conf. Comput. Vis. Image Process. Cham, Switzerland: Springer, 2020, pp. 115–126.

[15] Z. Sun, H. Chen, F. Shi, L. Wang, W. Zhu, D. Xiang, C. Yan, L. Li, and X. Chen, "An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images," Sci. Rep., vol. 6, no. 1, p. 21739, Feb. 2016.

[16] K. K. W. Cheng, "Macular vessel density, branching complexity and foveal avascular zone size in normal tension glaucoma," Sci. Rep., vol. 11, no. 1, pp. 1–9, Dec. 2021.

[17] Z. S. Y. Wong, J. Zhou, and Q. Zhang, "Artificial intelligence for infectious disease big data analytics," Infection, Disease Health, vol. 24, no. 1, pp. 44 48, Feb. 2019.

[18] X. Li, X. Feng, X. Sun, N. Hou, F. Han, and Y. Liu, "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019," Front. Aging Neurosci., vol. 14, 2022, Art. no. 937486.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

[19] G. P. Wormser, "The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: Clinical practice guidelines by the infectious diseases society of America," Clin. Infectious Diseases, vol. 43, no. 9, pp. 1089–1134, Nov. 2006, doi: 10.1086/508667.

[20] M.-N. Delyfer, "Adapted surgical procedure for Argus II retinal implantation: Feasibility, safety, efficiency, and postoperative anatomic findings," Ophthalmol. Retina, vol. 2, no. 4, pp. 276–287, Apr. 2018.

[21] A. K. Shukla, R. K. Pandey, and R. B. Pachori, "A fractional filter based efficient algorithm for retinal blood vessel segmentation," Biomed. Signal Process. Control, vol. 59, May 2020, Art. no. 101883.

[22] G. J. Chader, J. Weiland, and M. S. Humayun, "Artificial vision: Needs, functioning, and testing of a retinal electronic prosthesis," Prog. Brain Res., vol. 175, no. 9, pp. 317–332, 2009.

[23] A. Horsager, R. J. Greenberg, and I. Fine, "Spatiotemporal interactions in retinal prosthesis subjects," Investigative Opthalmol. Vis. Sci., vol. 51, no. 2, p. 1223, Feb. 2010.

[24] P. Opěla, I. Schindler, P. Kawulok, R. Kawulok, S. Rusz, and M. Sauer, "Shallow and deep learning of an artificial neural network model describing a hot flow stress evolution: A comparative study," Mater. Des., vol. 220, Aug. 2022, Art. no. 110880.

[25] R. Collu, E. J. Earley, M. Barbaro, and M. Ortiz-Catalan, "Non-rectangular neurostimulation waveforms elicit varied sensation quality and perceptive f ields on the hand," Sci. Rep., vol. 13, no. 1, p. 1588, Jan. 2023.