# Harnessing the power of vision transformers for enhanced OCT image classification

**Elena-Anca PARASCHIV[1,2], Alina-Elena SULTANA[2]**

[1] National Institute for Research & Development in Informatics – ICI Bucharest, Romania

[2] Faculty of Electronics, Telecommunications and Information Technology, National University of Science and Technology Politehnica Bucharest

elena.paraschiv@ici.ro, alina_elena.sultana@upb.ro

**Abstract:** The rising prevalence of eye disorders has raised concerns, emphasizing the need to accelerate the detection of retinal diseases. Early and accurate classification of these conditions is crucial for timely diagnosis and effective treatment in order to address critical situations. The recent advancements in retinal imaging have enhanced the diagnosis and management of Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME) or Drusen and the deep learning-based applications on Optical Coherence Tomography (OCT) images have further revolutionized the field by enabling automated, precise, and efficient disease classification, paving the way for earlier interventions and improved patient outcomes. This study investigates the use of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) for automated retinal disease classification. Three models were implemented: ViT, DeepViT, and a hybrid model combining ResNet50 with ViT, trained and evaluated on a publicly available OCT dataset. The hybrid model achieved the highest accuracy of 99.97%, thanks to its ability to capture both local and global features. This study underscores the potential of ViTs in medical image analysis and their integration with CNNs to develop accurate, robust, and scalable diagnostic tools, showing great promise for clinical applications.

**Keywords:** Vision Transformers (ViTs), OCT, Image Classification, Convolutional Neural Networks (CNNs), retina.

# Utilizarea potențialului Vision Transformers pentru clasificarea îmbunătățită a imaginilor OCT

**Abstract:** Prevalența tot mai mare a tulburărilor oculare a stârnit preocupări, subliniind necesitatea accelerării detectării bolilor retiniene. Clasificarea timpurie și precisă a acestor afecțiuni este crucială pentru diagnosticarea la timp și tratamentul eficace, pentru a aborda situațiile critice. Progresele recente în imagistica retiniană au îmbunătățit diagnosticarea și managementul Neovascularizației Coroidale (CNV), Edemului Macular Diabetic (DME) sau acumulărilor de tip Drusen, iar aplicațiile bazate pe învățarea profundă pe imagini de Tomografie în Coerență Optică (OCT) revoluționează în continuare domeniul, permițând clasificarea automată, precisă și eficientă a bolilor, ceea ce contribuie la intervenții mai timpurii și la îmbunătățirea rezultatelor pentru pacienți. Acest studiu investighează utilizarea Vision Transformers (ViTs) și a Rețelelor Neuronale Convoluționale (CNNs) pentru clasificarea automată a bolilor retiniene. Au fost implementate trei modele: ViT, DeepViT și un model hibrid care combină ResNet50 cu ViT, antrenate și evaluate pe un set de date OCT disponibil public. Modelul hibrid a atins cea mai mare acuratețe de 99.97%, datorită capacității sale de a capta atât caracteristici locale, cât și globale. Acest studiu subliniază potențialul ViTs în analiza imaginilor medicale și integrarea lor cu CNNs pentru dezvoltarea unor metode de diagnostic precise, robuste și scalabile, arătând un mare potențial pentru aplicații clinice.

**Cuvinte cheie:** Vision Transformers, OCT, clasificarea imaginilor, CNN, retina.

## 1. Introduction

In recent years, the realm of medical image analysis has been revolutionized by the onset of advanced deep learning (DL) techniques. These innovations promise to improve diagnostic accuracy, reduce human error, and improve patient outcomes (Puiu et al., 2021). Among these advancements, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have emerged as a novel technology, bringing a new level of precision and efficiency to image classification tasks traditionally dominated by Convolutional Neural Networks (CNNs). Originally designed for natural language processing (NLP), the Transformer architecture's ability to capture long-range

https://doi.org/10.33436/v34i2y202408

dependencies and contextual relationships has now found compelling applications in the realm of computer vision.

The application of ViTs in medical imaging is particularly promising. Their attention-based mechanisms allow for the detailed analysis of complex anatomical structures, surpassing the capabilities of conventional methods. This paradigm shift is not just theoretical; it has real-world implications for the early detection and treatment of diseases (Dai & Gao, 2021; Gao et al., 2021; Mondal et al., 2021). In addition to this, in fields like ophthalmology, where timely diagnosis can prevent severe vision loss, the potential impact of ViTs can be profound.

Optical coherence tomography (OCT) stands out as a revolutionary imaging technique in ophthalmology, providing high-resolution, cross-sectional images of the retina. These images have become indispensable in the detailed visualization and diagnosis of various retinal conditions, significantly enhancing the ability to detect and monitor diseases at an early stage. Among the numerous retinal diseases, choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen are particularly prevalent and visually debilitating. CNV is associated with abnormal blood vessel growth beneath the retina, leading to vision loss if untreated (Shah, n.d.). DME, a complication of diabetic retinopathy, involves the accumulation of fluid in the macula through microaneurysms that form in the blood vessels, impairing the central vision (Cleveland Clinic, n.d.). Drusen, characterized by yellow deposits under the retina, is a common feature of age-related macular degeneration (AMD) and can lead to progressive vision impairment (What Are Drusen, 2019).

Despite the clinical importance of OCT, interpreting these images remains a complex task requiring significant expertise. Ophthalmologists must discern subtle changes in retinal structure, a process that is both time-consuming and prone to subjective variability. The high volume of images that need to be reviewed exacerbates these challenges, increasing the risk of inconsistencies and missed diagnoses.

To overcome these challenges, there is growing interest in integrating DL techniques into the analysis of OCT images. ViTs, with their superior ability to model global context and intricate details, present a compelling solution. By automating the classification of retinal diseases, ViTs can enhance diagnostic accuracy and consistency, providing reliable support to ophthalmologists.

Several algorithms that leverage the strengths of ViTs and explore hybrid models that combine the capabilities of CNNs and ViTs are proposed. This approach aims to enhance the accuracy and robustness of retinal disease classification, ultimately contributing to better diagnostic tools for ophthalmologists. The contributions are threefold:

- Applying ViT-based algorithms tailored for CNV, DME, and drusen classification in OCT images;
- Proposal of a hybrid model that integrates the strengths of CNNs and ViTs;
- Extensive experiments and comparative analyses to validate the effectiveness of the methods.

This paper harnesses the power of ViTs for the classification of retinal diseases from OCT images and it is organized as follows: Section 2 presents a comprehensive review of the most recent advancements in retinal disease classification using ViTs. Section 3 focuses on the proposed approach, detailing the methods employed, the results obtained, and a discussion of the main findings and the final section provides the conclusion, summarizing the key outcomes and implications of the study.

## 2. State of the art

In the last decade, the applications of DL techniques to the classification of retinal diseases using OCT images have gained significant traction. While numerous methods leveraging CNNs have been extensively studied and applied for the detection of retinal diseases (Choudhary et al., 2023; Elkholy & Marzouk, 2024; Nawaz et al., 2023), the focus has increasingly shifted towards ViTs due to their promising performance and ability to capture complex patterns in medical images. The following state of the art was chosen to include papers that focus on the same dataset

that was used in the present paper, ensuring a consistent basis for comparison. Various studies have explored different architectures and methodologies to enhance the accuracy and reliability of automated diagnostic systems. This section provides a summary of several notable works in this domain, highlighting their methodologies, results, and contributions to the field.

A significant advancement in this domain is presented in the study performed by Jingzhen He (He et al., 2023). This research proposes the Swin-Poly Transformer network, which utilizes a shifting window partition approach to connect neighbouring non-overlapping windows from the previous layer, thereby effectively modelling multi-scale features. Additionally, the Swin-Poly Transformer refines cross-entropy by adjusting the significance of polynomial bases, enhancing classification performance. This method not only achieves high accuracy, but also generates confidence score maps, aiding medical practitioners in understanding the model's decision-making process. The results indicate superior performance, with an accuracy of 99.80% and an AUC of 99.99% on the OCT2017 and OCT-C8 datasets, surpassing the performance of traditional CNNs.

Another study (Hemalakshmi et al., 2024) proposed a hybrid model named SqueezeNet-ViT (SViT), which combines the capabilities of SqueezeNet and ViTs for retinal disease classification using OCT images. This hybrid approach leverages both local feature extraction and global contextual information, resulting in a more accurate and computationally efficient model. The SViT model was evaluated on the OCT2017 dataset for both binary and multiclass classification tasks, achieving an overall classification accuracy of 99.90%.

A study focused on the application of ViT for retinal diseases diagnosis using OCT images (Zhou et al., 2023). ViTs, introduced in 2020, utilize a transformer-based architecture entirely for feature extraction, differing from traditional CNNs. They employed a ViT model to classify OCT images into normal, CNV, Drusen, and DME categories. The results showed that the model achieved an accuracy of 95.76%, sensitivity of 95.77%, and specificity of 98.59%. These metrics indicate that ViT can outperform traditional CNN models in the classification of retinal diseases, offering an effective tool for early diagnosis and treatment planning.

The advancements in DL and ViT have significantly improved the classification of retinal diseases using OCT images. Studies have proved the efficacy of various architectures, including interpretable transformers, hybrid models, and standalone ViTs. These models have achieved remarkable accuracy and reliability, highlighting their potential to enhance clinical decision-making and patient care. The present paper further builds on this progress by implementing and evaluating the ViT, DeepViT (Zhou et al., 2021), and a hybrid CNN-ViT model, achieving good performance metrics and offering insights into their practical applicability in clinical settings. Future research should focus on validating these models on diverse datasets, optimizing computational efficiency, and integrating them into clinical workflows to fully realize their benefits in ophthalmology.

# 3. The proposed approach

## 3.1. Dataset

The dataset utilized in this study is sourced from a publicly available repository (Kermany et al., 2018) on Kaggle, a prominent data science platform that provides robust tools for researchers and developers. It comprises OCT images of the retina, including images of healthy retinas as well as those affected by the three distinct retinal diseases: CNV, DME and Drusen. OCT is an advanced imaging technique that offers high-resolution cross-sectional views of the human retina. Each image in the dataset has been meticulously graded by ophthalmologists, ensuring high-quality annotations. The use of light reflection in the capturing process of OCT images contributes to their exceptional quality, free from biases towards any specific disease category. The images are systematically categorized into four directories corresponding to the labels: CNV, DME, Drusen, and Normal. This comprehensive dataset provides a robust foundation for the proposed work, enabling the development of models that can accurately classify retinal diseases based on OCT images.

The dataset is organized into training, validation and testing subsets to facilitate the development and evaluation of DL models. Specifically, the training set consists of 83,484 images, the validation set contains 32 images, and the test set includes 968 images, the distribution being illustrated in Figure 1.

To ensure the quality and consistency of the images, several pre-processing steps were undertaken, including:

- *Resizing* to match the requirements of the neural networks. All images were resized to 224x224 pixels using bilinear interpolation. This dimension was chosen to balance between computational efficiency and the retention of important image details;
- *Normalization* to facilitate faster convergence during model training. Pixel values were scaled from the original [0, 255] range to [0, 1] by dividing each pixel value by 255. This normalization ensures that the input values are consistent and helps in stabilizing the training process;
- *Data augmentation* to increase the diversity of the training dataset and prevent overfitting by simulating variations seen in real-world data. The techniques that were used include rotation, flipping, scaling, translation.

These pre-processing steps enhance the model's robustness and generalizability. Despite the strengths of this dataset, several potential limitations and challenges must be addressed, such as class imbalance, variability in image quality, and the need for extensive pre-processing techniques.

Moreover, ethical considerations are paramount when using publicly available medical datasets, particularly regarding patient privacy and data security. The dataset from Kaggle is anonymized, ensuring that no personally identifiable information is included.

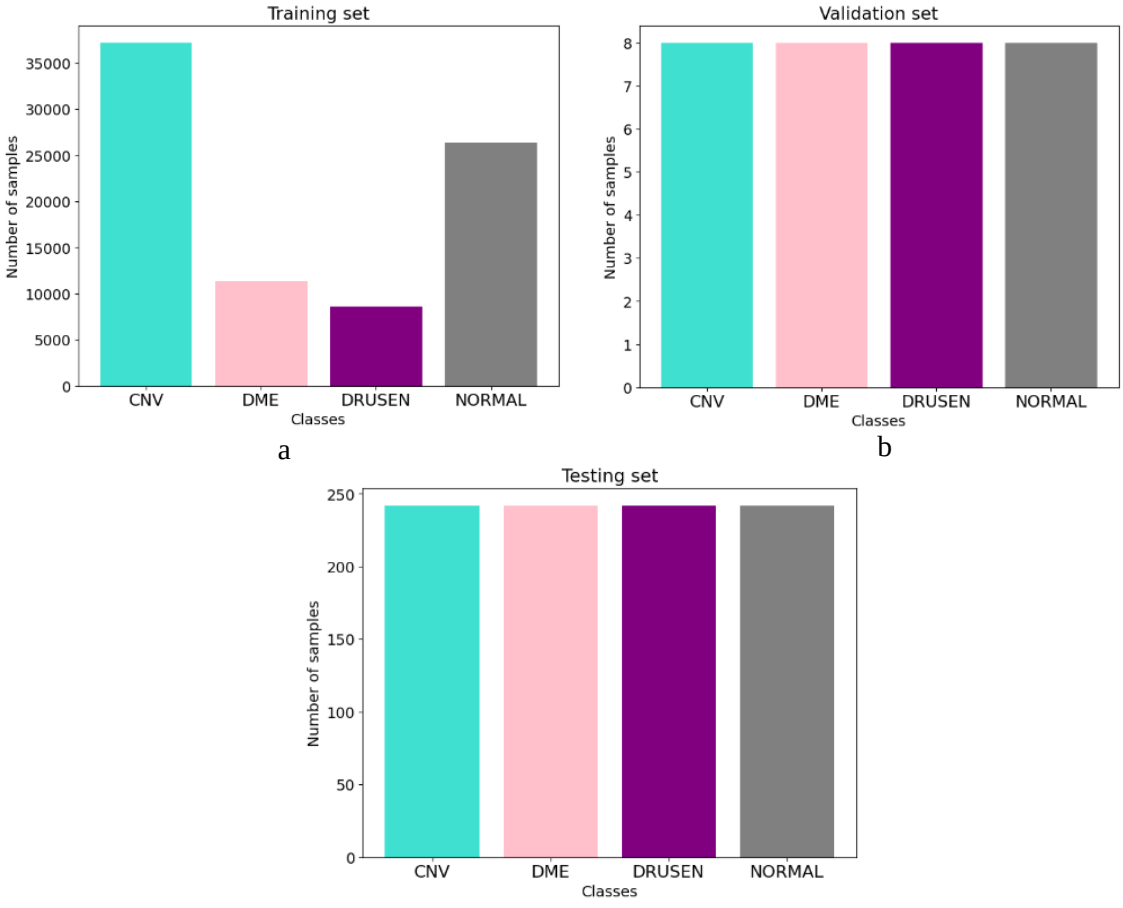



**Figure 1.** The distribution of the (a) training, (b) validation and (c) testing datasets

The actual numbers of images per class for training, validation and testing are presented in the following table (see Table 1).

**Table 1.** Number of images for each class for training, validation and testing

| Set | CNV | DME | Drusen | Normal |
|---|---|---|---|---|
| Training | 37,205 | 11,348 | 8,616 | 26,315 |
| Validation | 8 | 8 | 8 | 8 |
| Testing | 242 | 242 | 242 | 242 |

A representative image from each class has been selected for presentation to facilitate clear visualization of the differences between each class (see Figure 2). Blue arrows highlight the retinal structures associated with each disease.

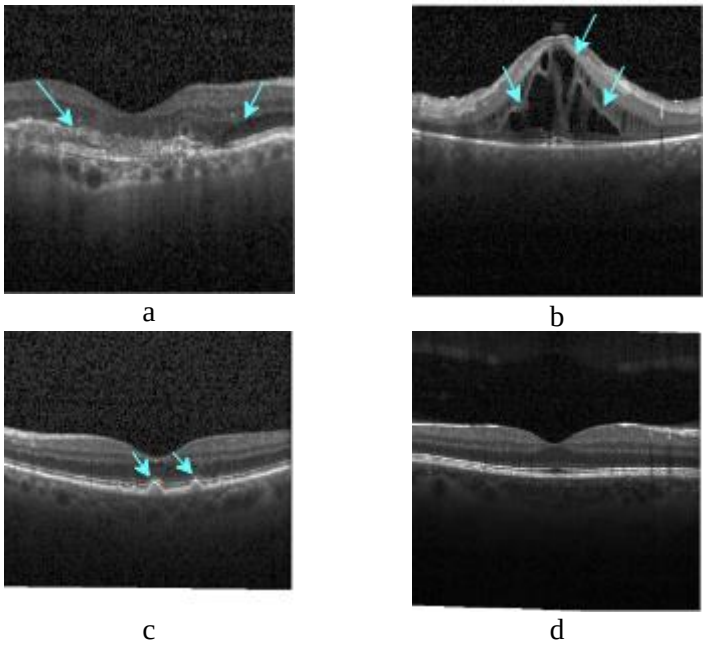



**Figure 2.** A representative image for: (a) CNV; (b) DME; (c) Drusen and (d) Normal (Kermany et al., 2018)

## 3.2. Methods

This chapter delineates the methodologies employed in the development and evaluation of DL models for the classification of retinal diseases using OCT images. The focus is on three approaches: ViT, DeepViT, and a hybrid combination of CNNs and ViT. Each technique presents distinct advantages and addresses specific challenges in the classification of retinal diseases. To provide a comprehensive understanding of ViTs, it is essential to first explore the foundational Transformer architecture upon which ViTs are built (Henry et al., n.d.).

The Transformer architecture, introduced by Vaswani et al. (2017), revolutionized NLP with its attention mechanisms that efficiently handle long-range dependencies, outperforming recurrent neural networks in tasks like text translation, natural language generation, and speech recognition. Central to its design, the architecture employs self-attention mechanisms and point-wise feed-forward networks (FFN) to process sequences in parallel, capturing intricate inter-token relationships without sequential constraints. Self-attention, the core of this model, uses queries, keys and values to focus selectively on different segments of the input sequence, thus enabling contextually rich interpretations. This is enhanced by multi-head attention which concurrently explores various aspects of the sequence, combining these perspectives to form a comprehensive representation (Floroiu & Timisică, 2024).

Additionally, Transformers utilize positional encodings to imbue sequence order awareness, essential for maintaining token positional context. Since its debut, the Transformer has inspired

numerous variants like the Bidirectional Encoder Representations from Transformers - BERT and Generative Pre-trained Transformer - GPT, significantly advancing NLP by enabling more nuanced understanding and generation of text.

Unlike CNNs, which aggregate global information by stacking multiple convolutional layers (e.g., 3 × 3), ViTs leverage the self-attention mechanism to capture spatial patterns and non-local dependencies. Several notable transformer-based vision models have been developed, including the first version of ViT, DEtection TRansformer (DETR) (Carion et al., 2020), which is an encoder-decoder-based Transformer architecture that simplifies object detection, or Swin-Transformer (Liu et al., 2021), which is based on a hierarchical Transformer computed with shifted windows. It restricts self-attention to local windows while allowing cross-window connections, effectively handling different visual scales (Floroiu et al., 2024).

The input images in ViT are divided into a sequence of non-overlapping patches (Figure 3), with each patch represented as a vector. Positional information is incorporated into these vectors and fed into the transformer's encoder, which includes multi-head self-attention, layer normalization, and an FFN. Understanding the variations and advancements in these transformer-based models can help in the improvement of their application and effectiveness in the domain of medical image analysis, particularly for the classification of retinal diseases using OCT images.

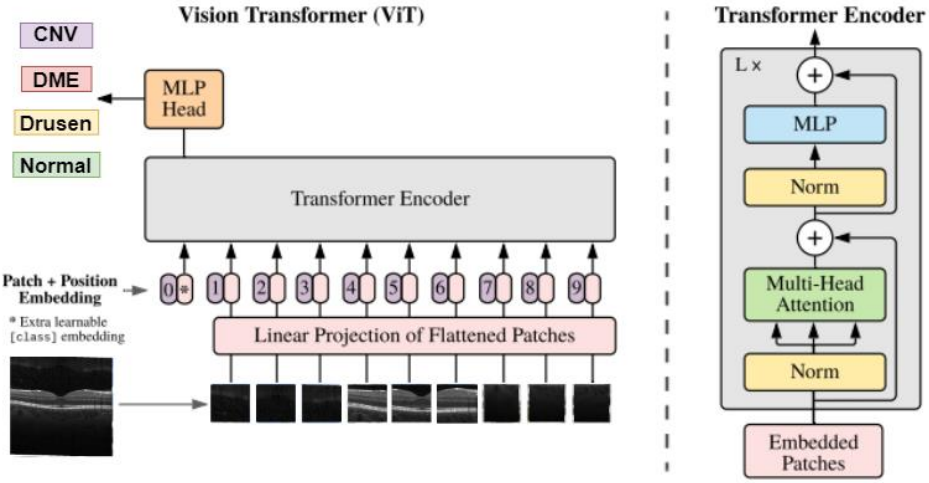



**Figure 3.** ViT (adapted from Dosovitskiy et al., 2020)

An interesting extension of ViT is DeepViT which is designed to enhance the model's capability to capture deeper and more complex patterns in image data. While ViT partitions an image into patches and processes these patches through a series of transformer encoders to capture global dependencies and spatial relationships, DeepViT incorporates additional layers and optimization techniques to improve feature extraction and representation. DeepViT addresses some of the limitations of ViT, such as its performance dependency on large-scale datasets, by introducing more sophisticated mechanisms for hierarchical feature learning and regularization. This makes DeepViT more robust in learning intricate patterns and relationships within the data, thereby offering improved accuracy and generalization in the classification of retinal diseases using OCT images. The enhancements in DeepViT enable it to better handle the complexities and nuances present in medical imaging, leading to more precise and reliable diagnostic outcomes.

In this study, the vit_pytorch library is utilized to implement Vision Transformer (ViT) and DeepViT models. These models were pretrained on the ImageNet dataset, allowing to leverage transfer learning for efficient and effective feature extraction. The pretrained models were fine-tuned on the publicly available OCT dataset to classify CNV, DME and Drusen. This section further presents the models that were implemented and evaluated: ViT, DeepViT and the hybrid CNN-ViT model. The hybrid model combined ResNet50, a CNN known for its local feature extraction capabilities, with ViT, which proved to excel at capturing global features.

### 3.2.1. ViT

The key parameters for the model include: dimension – 64 (the size of the embedding space for each image patch, determining how much information each patch can represent); depth - 8 transformer blocks (the number of layers in the model, with each layer learning progressively more complex features from the input data); heads - 8 attention heads (the number of parallel attention mechanisms, allowing the model to focus on different parts of the data simultaneously and capture diverse relationships); MLP dimension – 128 (the size of the hidden layers in the FFN within each transformer block, enabling the model to learn complex transformations and representations of the input data).

The ViT model processes input images by dividing them into non-overlapping patches, each of which is linearly embedded. These patches are then fed into a series of transformer encoders that apply multi-head self-attention and feed-forward layers to capture both local and global dependencies in the data.

To address the class imbalance in the training data, class weights were computed based on the frequency of each class. These weights were used to create a weighted random sampler, ensuring that each class was adequately represented during training and mitigating the risk of the model being biased towards the majority class.

The training parameters were set as follows: batch size – 32; epochs – 20; learning rate - 0.0003. During training, images were loaded in batches using PyTorch's data loader, with the training set utilizing the weighted random sampler. The model was trained using the Adam optimizer and the training loop involved feeding batches of images through the model, computing the loss using negative log likelihood, backpropagating the gradients, and updating the model weights.

### 3.2.2. DeepViT

The implementation of the DeepViT model follows a similar structure to the ViT model, with enhancements designed to improve the model's depth and representational capacity. The DeepViT model incorporates additional transformer blocks compared to the standard ViT model. This increased depth allows the model to learn more complex and hierarchical features from the input data. Each additional transformer block consists of multi-head self-attention layers followed by FFN, enabling the model to perform multiple levels of transformation and feature extraction.

DeepViT also includes improvements to the self-attention mechanism to better capture intricate dependencies and relationships within the data, which may involve more sophisticated attention head configurations or advanced normalization techniques to stabilize and improve learning. The FFNs within each transformer block of the DeepViT model are more extensive, with increased MLP dimensions. This allows the model to process the output of the attention mechanism with greater complexity, leading to more refined feature representations.

### 3.2.3. Hybrid CNN-ViT

In addition to the ViT and DeepViT models, a hybrid model was also proposed that combines the strengths of CNNs and ViTs. This hybrid approach leverages the feature extraction capabilities of CNNs with the global context capturing abilities of ViTs to improve the classification of retinal diseases using OCT images.

*Model Architecture:* The hybrid model consists of two main components: a pre-trained ResNet50 CNN and a ViT. The architecture (Figure 4) is designed to extract and refine features through the following steps:

- Feature extraction with ResNet50: The ResNet50 model, pre-trained on ImageNet, is used as the initial feature extractor. The final fully connected layer of ResNet50 is removed, and the output feature map is passed to the next stage;
- Transformation to ViT input: The extracted features from ResNet50 are transformed to

match the input dimensions required by the ViT model. A linear layer is used to adjust the dimensionality of the feature map;

- ViT processing: The transformed feature map is fed into the ViT, which consists of multiple transformer blocks. Each block includes multi-head self-attention mechanisms and FFNs, allowing the model to capture complex and global dependencies;
- Classification head: Finally, a fully connected layer is added to the output of the ViT to perform the classification into the desired number of classes (four in this case, corresponding to the retinal diseases categories).

The model was trained using the following parameters: batch size – 32; epochs – 10; learning rate - 0.001.

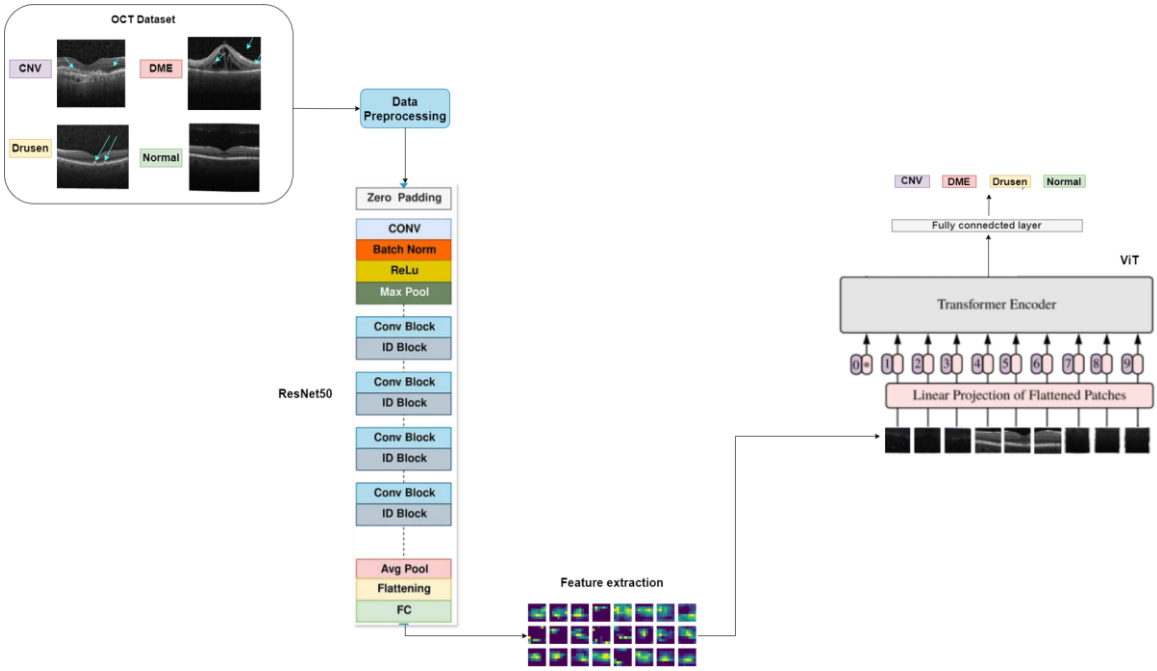


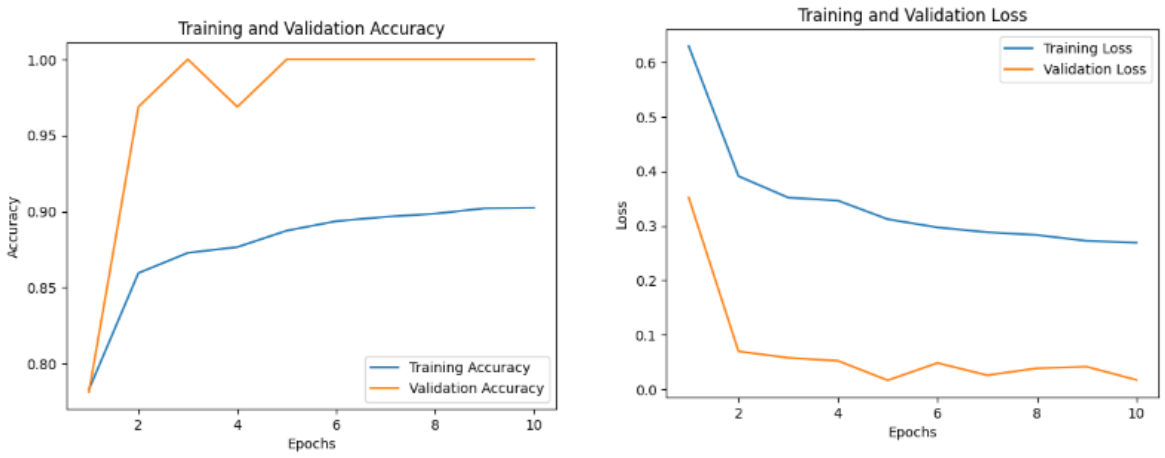

**Figure 4.** The Hybrid CNN-ViT model approach (adapted from Dosovitskiy et al., 2020)

## 3.3. Results

In this section, the outcomes of the experiments are presented, each approach being carefully designed and implemented to leverage the strengths of transformer architectures and CNNs, addressing the unique challenges posed by medical image analysis. The performances for each model are presented in Table 2, as well as a comprehensive comparison of their effectiveness in accurately classifying retinal diseases. Figure 5 also shows the training and validation loss and accuracy of the hybrid model.

**Table 2.** The performance of the models

| Model | Accuracy | Precision | Recall | F1-score | Time (s) |
|---|---|---|---|---|---|
| ViT | 96.80% | 96.80% | 96.80% | 0.96 | 19627.86 |
| DeepViT | 89.98% | 90.03% | 89.98% | 0.90 | 24632.96 |
| Hybrid model ResNet50-ViT | 99.97% | 99.10% | 99.97% | 0.99 | 16220 |




a b

**Figure 5.** (a) The training and validation loss and (b) accuracy for the hybrid model

The confusion matrices for the models are described in Figure 6 as they provide a detailed evaluation of the models' performances in classifying the four categories of retinal diseases.

**ViT model:** The confusion matrix for ViT is shown in Figure 6a. The ViT model achieved an accuracy of 96.80%, with precision, recall, and F1-score all at 96.80%. The matrix indicates that the model performs well across all classes, with a few misclassifications observed primarily in the Drusen and Normal categories. Specifically, 7 Drusen cases were misclassified as CNV, and 7 Normal cases were misclassified as Drusen. This suggests that while the ViT model is highly effective, there is some room for improvement in distinguishing between these categories.

**DeepViT model:** The confusion matrix for the DeepViT model (Figure 6b) illustrates a lower overall performance compared to the ViT model, having an accuracy of 89.98%. Precision, recall, and F1-score are around 90%. The matrix shows that the DeepViT model has more difficulty accurately classifying certain categories. For instance, there are notable misclassifications of Drusen as CNV (19 cases) and Normal (17 cases), and DME as CNV (14 cases) and Normal (7 cases). These results indicate that while the DeepViT model captures some complex patterns, it struggles more with inter-class differentiation compared to the ViT model.

**Hybrid Model (ResNet50-ViT):** The hybrid model (Figure 6c), which combines ResNet50 with ViT, achieved the highest performance metrics among the three models. It recorded an accuracy of 99.97%, with precision, recall, and F1-score all around 99%. The confusion matrix reveals that the hybrid model nearly perfectly classifies all categories, with minimal misclassifications. Only 8 Drusen cases were incorrectly classified as CNV. This highlights the hybrid model's ability to leverage both CNN and transformer architectures to improve classification accuracy significantly.

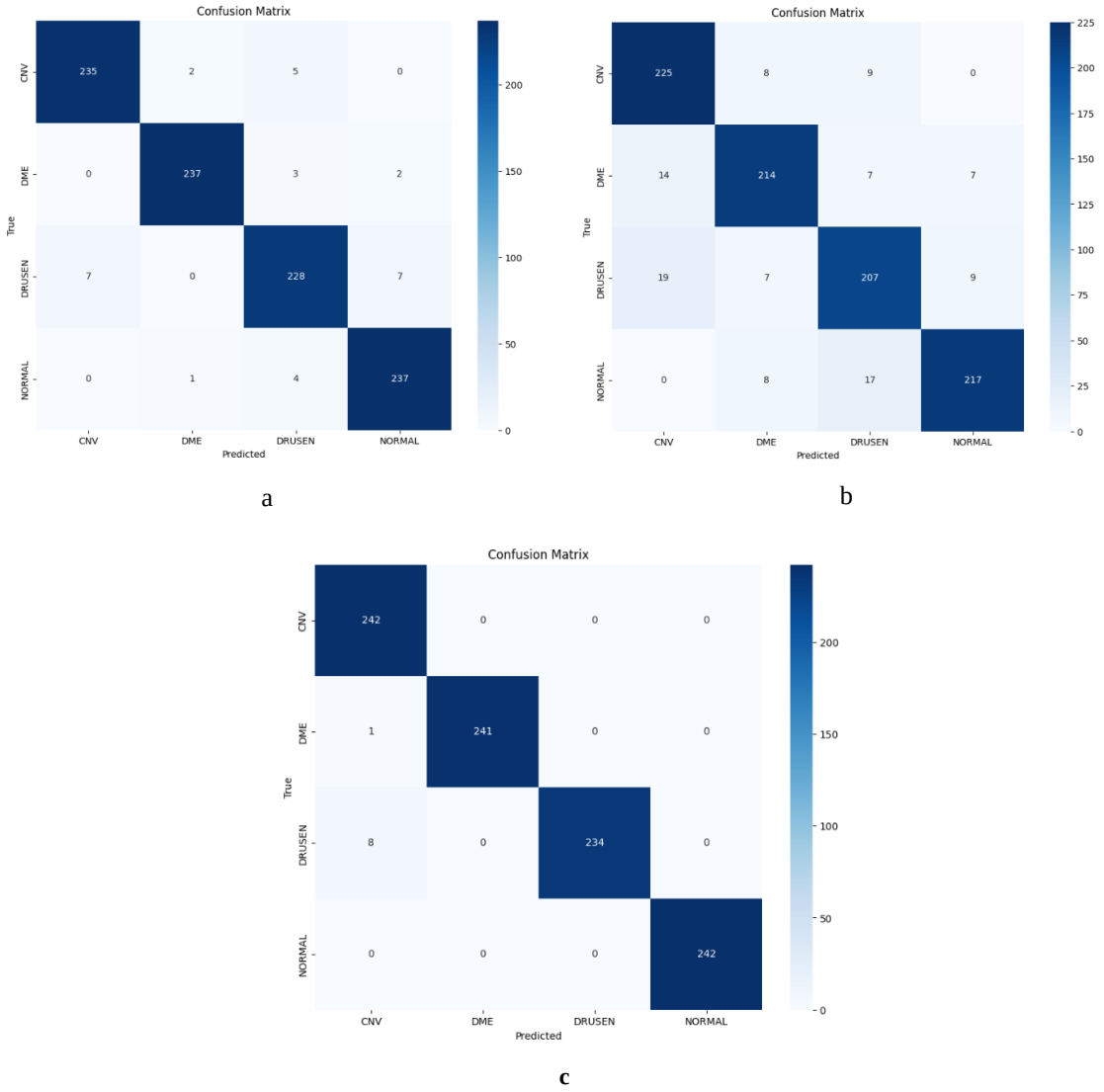



**Figure 6.** (a) The confusion matrices for: ViT; (b) DeepViT; (c) Hybrid model

Based on the above-mentioned results, the hybrid model significantly outperforms both the ViT and DeepViT models in all key performance metrics. The ViT model, while performing admirably, still shows some misclassifications, particularly between Drusen and Normal categories. The DeepViT model, despite its complexity and depth, does not perform as well as the ViT model, indicating potential overfitting or insufficient learning of inter-class features. The hybrid model's integration of CNN and ViT architectures enables it to capture both local and global features more effectively, leading to near-perfect classification accuracy and demonstrating its robustness and reliability in clinical settings.

## 4. Discussions

In this section, the performance of the algorithms is compared. Each model was evaluated based on its ability to classify retinal diseases using OCT images, with key metrics including accuracy, precision, recall, F1-score, and training time.

The ViT model demonstrated robust performance throughout the training process. Over the course of 20 epochs, the ViT model showed a significant reduction in average test loss from 0.8046 to 0.1276. Correspondingly, the accuracy increased from 70.35% in the first epoch to 96.80% by the 20th epoch. Precision, recall, and F1-score metrics were all consistently high at 96.80%, indicating balanced and reliable performance across all classes.

The DeepViT model also exhibited substantial improvements over the 20 epochs of training. Starting with a higher initial loss of 1.3724, the model's average test loss decreased to 0.3531 by the final epoch. The accuracy improved significantly from 52.69% in the first epoch to 89.98% in the 20th epoch. Precision, recall, and F1-score were all around 90%, demonstrating the model's effectiveness in classifying retinal diseases.

The hybrid model, which combines the strengths of CNNs and ViTs, showed remarkable performance improvements over a shorter training period of 10 epochs. The training accuracy steadily increased, reaching 90.10% by the final epoch, and the validation accuracy consistently achieved almost 99% from epoch 2 onwards. The final test accuracy was 99.38%, with precision, recall, and F1-score all near 99%. The hybrid model's ability to leverage both CNN's feature extraction and ViT's contextual learning capabilities resulted in superior performance metrics. Additionally, the classification report (Figure 7) for the hybrid model indicated near-perfect classification across all retinal disease's categories.

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| CNV | 0.98 | 1.00 | 0.99 | 242 |
| DME | 1.00 | 0.99 | 0.99 | 242 |
| DRUSEN | 0.99 | 1.00 | 0.99 | 242 |
| NORMAL | 1.00 | 1.00 | 1.00 | 242 |
| accuracy | | | 0.99 | 968 |
| macro avg | 0.99 | 0.99 | 0.99 | 968 |
| weighted avg | 0.99 | 0.99 | 0.99 | 968 |

**Figure 7.** The classification report for the hybrid model

The models were trained on the same GPU setup provided by Kaggle, specifically utilizing the T4x2 configuration and to what concerns the training time, the total training time for the ViT model was approximately 19627.86 seconds. The DeepViT model, due to its increased complexity and depth, required a longer training time of 24632.96 seconds. Despite its superior performance, the hybrid model was notably efficient, requiring a shorter training period of just 10 epochs in approximately 16220 seconds, highlighting the hybrid model's efficiency in achieving high performance without the need for prolonged training.

Moreover, the following table shows the performance of the methods that dived into the classification of retinal diseases using ViT-based architectures, compared to the proposed hybrid model that used ResNet50 and ViT.

**Table 3.** Performance of other methods compared to the proposed hybrid model

| **Paper** | **Method** | **Accuracy (%)** |
|---|---|---|
| He, J., 2023 | Swin-Poly Transformer | 99.80 |
| Hemalakshmi, G.R., 2024 | SqueezeNet + ViT | 99.90 |
| Zhou, Z., 2023 | ViT | 95.77 |
| Proposed hybrid model | ResNet50 + ViT | 99.97 |

In comparison to the study by Hemalakshmi, G.R. (2024), which employed SqueezeNet alongside ViT for OCT image classification, this study utilized ResNet50 as the CNN component of the hybrid model. ResNet50 was selected due to its deeper architecture and superior performance in feature extraction tasks, as evidenced by its widespread use in various computer vision applications. The primary shortcomings addressed in this work include:

- Feature extraction: SqueezeNet, while efficient, has a lighter architecture which may not capture as complex features as ResNet50. By incorporating ResNet50, our model benefits from a deeper network that can extract more nuanced features from OCT images;
- Classification accuracy: The experimental results from this study demonstrate a significant improvement in classification accuracy, with the hybrid model achieving 99.97% accuracy compared to the previously reported results using SqueezeNet.

Across the different models, a notable pattern in misclassification was observed, particularly concerning the Drusen category. Drusen was frequently misclassified as either CNV or Normal. This trend was evident in both the ViT and DeepViT models. For the ViT model, there were 7 instances of Drusen misclassified as CNV and 7 as Normal. The DeepViT model showed a higher rate of misclassification with 19 Drusen cases misclassified as CNV and 17 as Normal. This misclassification issue might be attributed to the subtle and overlapping features between Drusen and the other categories, which make it challenging for the models to distinguish them accurately. Drusen, being a type of extracellular deposit beneath the retina, shares visual similarities with certain characteristics of CNV and Normal, which can lead to confusion during classification. The hybrid model demonstrated an improved performance in this regard, with only 8 Drusen cases misclassified, underscoring the benefit of combining CNN's feature extraction capabilities with the global attention mechanism of ViTs.

Addressing these misclassifications would require further refinement of the models, potentially incorporating additional domain-specific features or enhancing the training dataset to better capture the distinguishing characteristics of Drusen.

Despite the promising results, this study has several limitations. First, the models were trained and evaluated on a single dataset, which may not fully represent the diversity and complexity of retinal diseases encountered in clinical practice. Consequently, the generalizability of the findings to other datasets and real-world scenarios might be limited. Second, the computational resources required for training transformer-based models, particularly the DeepViT, are substantial. This poses a challenge for implementation in resource-constrained environments, such as smaller medical facilities or regions with limited access to high-performance computing infrastructure. Third, while the hybrid model demonstrated superior performance, the integration of CNN and ViT architectures adds complexity to the model design and training process. This complexity might hinder the adoption and scalability of the approach in practice. Lastly, this study did not extensively explore hyperparameter tuning or the impact of different data augmentation techniques, which could potentially further enhance the performance of the models. Future work should address these limitations by evaluating the models on more diverse datasets, optimizing computational efficiency, simplifying model architectures, and exploring advanced training techniques.

## 5. Conclusions

This study explored the effectiveness of three distinct DL approaches – ViT, DeepViT, and a hybrid model combining CNNs with ViT – for the classification of retinal diseases using OCT images. The findings demonstrate that each model has unique strengths and limitations, contributing valuable insights into their applicability for medical image analysis.

The ViT model exhibited strong performance with a final accuracy of 96.80%, showcasing its capability in handling complex visual tasks. However, the DeepViT model, despite its improved depth and complexity, achieved a slightly lower accuracy of 89.98%, reflecting the trade-off between model complexity and performance. The hybrid model outperformed both, achieving a near-perfect test accuracy of 99.38%, illustrating the efficacy of combining CNN's feature extraction capabilities with ViT's contextual learning.

The comparative analysis revealed that the hybrid model not only excelled in accuracy but also in precision, recall and F1-score, all around 99%. This model also demonstrated exceptional generalization with consistent validation accuracy of almost 100% from the second epoch onward. Despite the superior performance, the hybrid model was more efficient, requiring fewer epochs and less time to train, highlighting its practical benefits for clinical applications.

However, the study also identified several limitations. The reliance on a single dataset may limit the generalizability of the findings, and the substantial computational resources required for training, particularly for transformer-based models, pose challenges for widespread adoption. Additionally, the complexity of integrating CNN and ViT architectures might hinder scalability.

In conclusion, the research underscores the potential of advanced DL models, particularly hybrid architectures, in enhancing the accuracy and reliability of retinal disease classification. Future work should focus on addressing the identified limitations by evaluating the models on diverse datasets, optimizing computational efficiency, simplifying model designs, and exploring advanced training techniques. These efforts will be crucial in advancing the practical application of DL in medical image analysis, ultimately improving diagnostic accuracy and patient outcomes in ophthalmology.

# REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://arxiv.org/abs/2005.12872 [Accessed 10th May 2024].

Choudhary, A., Ahlawat, S., Urooj, S., Pathak, N., Lay-Ekuakille, A. & Sharma, N. (2023) A Deep Learning-Based Framework for Retinal Disease Classification. *Healthcare*. 11(2), 212. doi:10.3390/healthcare11020212.

Cleveland Clinic (n.d.) *What is Diabetes-Related Macular Edema (DME)?* https://my.clevelandclinic.org/health/diseases/24733-diabetes-related-macular-edema [Accessed 10th May 2024].

Dai, Y. and Gao, Y. (2021). TransMed: Transformers Advance Multi-modal Medical Image Classification. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://doi.org/10.48550/arxiv.2103.05940 [Accessed 15th May 2024].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. To be published in *Computer Vision and Pattern Recognition*. [Preprint] http://arxiv.org/abs/2010.11929 [Accessed 15th May 2024].

Elkholy, M. & Marzouk, M.A. (2024) Deep learning-based classification of eye diseases using Convolutional Neural Network for OCT images. *Frontiers in Computers Science.* 5, 1-12. doi:10.3389/fcomp.2023.1252295.

Floroiu, I., Floroiu, M., Niga, A. C. & Timisica D. (2024) Remote Access Trojans Detection Using Convolutional and Transformer-based Deep Learning Techniques. *Romanian Cyber Security Journal*. 6(1), 47-58. doi:10.54851/v6i1y202405.

Floroiu, I. & Timisică, D. (2024) A Heideggerian analysis of generative pretrained transformer models. *Romanian Journal of Information Technology and Automatic Control (Revista Română de Informatică și Automatică)*. 34(1), 13-22. doi: 10.33436/v34i1y202402.

Gao, X., Qian, Y. & Gao, A. (2021) COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models. To be published in *Image and Video Processing.* [Preprint] https://doi.org/10.48550/arXiv.2107.01682 [Accessed 16th May 2024].

He, J., Wang, J., Han, Z., Ma, J., Wang, C. & Qi, M. (2023) An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Scientific Reports*. 13, 3637. doi:10.1038/s41598-023-30853-z.

Hemalakshmi, G.R., Murugappan, M., Sikkandar, M. Y., Begum, S. & Prakash, N.B. (2024) Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images. *Neural Computing & Applications*. 36, 9171-9188. doi:10.1007/s00521-024-09564-7.

Henry, E.U., Emebo, O. & Omonhinmin, C.A. (n.d.) *Vision Transformers in Medical Imaging: A Review.* https://arxiv.org/pdf/2211.10043 [Accessed 30th May 2024].

Kermany, D., Zhang, K. & Goldbaum, M. (2018) Labeled Optical Coherence Tomography (OCT) and Chest X-ray Images for Classification. *Mendeley Data*. 2. doi: 10.17632/rscbjbr9sj.2.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://arxiv.org/abs/2103.14030 [Accessed 20th May 2024]

Mondal, A.K., Bhattacharjee, A., Singla, P. & Prathosh, A.P. (2022) xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography. In *IEEE Journal of Translational Engineering in Health and Medicine*. 10, 1–10. doi:10.1109/jtehm.2021.3134096.

Nawaz, A., Ali, T., Mustafa, G., Babar, M. & Qureshi, B. (2023) Multi-Class Retinal Diseases Detection Using Deep CNN With Minimal Memory Consumption. *IEEE Access*. 11, 56170-56180. doi:10.1109/ACCESS.2023.3281859.

Shah, M. (n.d.) *Choroidal neovascularization (CNV).* https://www.eyeclinic-karachi.com/choroidal-neovascularization-cnv/ [Accessed 10th May 2024].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017) Attention Is All You Need. To be published in *Computation and Language.* [Preprint] https://arxiv.org/abs/1706.03762 [Accessed 20th May 2024]

Porter, D. (2023) *What Are Drusen?* https://www.aao.org/eye-health/diseases/what-are-drusen [Accessed 15th May 2024]

Puiu, A., Vizitiu, A., Nita, C., Itu, L., Sharma, P., Comaniciu, D. (2021) Privacy-Preserving and Explainable AI for Cardiovascular Imaging. *Studies in Informatics and Control*. 30(2), 21-32. doi:10.24846/v30i2y202102.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q. & Feng, J. (2021) DeepViT: Towards Deeper Vision Transformer. To be published in *Computer Vision and Pattern Recognition.* [Preprint] https://arxiv.org/abs/2103.11886 [Accessed 20th May 2024].

Zhou, Z., Niu, C., Yu, H., Zhao, J., Wang, Y. & Dai, C. (2023) Diagnosis of retinal diseases using the vision transformer model based on optical coherence tomography images. In *Proceedings Volume SPIE-CLP Conference on Advanced Photonics 2022.* 1260102. doi:10.1117/12.2665918.

**Elena-Anca PARASCHIV** este cercetător științific la Departamentul „Ingineria Software și a Sistemelor Complexe” din cadrul Institutului Național de Cercetare-Dezvoltare în Informatică - ICI București și student-doctorand în cadrul Școlii Doctorale de Electronică, Telecomunicații și Tehnologia Informației, Universitatea Națională de Știință și Tehnologie POLITEHNICA București (UNSTPB). A absolvit Facultatea de Inginerie Medicală din cadrul UNSTPB și deține o diplomă de master în specializarea „Sisteme inteligente și vedere artificială” din cadrul Facultății de Electronică, Telecomunicații și Tehnologia Informației, UNSTP. Domeniile și subiectele sale de interes pentru activitatea de cercetare cuprind: aplicații bazate pe inteligența artificială, în special în domeniul medical (prelucrare și analiză de imagini și semnale medicale), aplicații de telemedicină și dezvoltarea de echipamente pentru asistență medicală.

**Elena-Anca PARASCHIV** is a Scientific Researcher in the ″Software Engineering and Complex Systems″ Department at the National Institute for Research and Development in Informatics - ICI Bucharest and a Ph.D. candidate in the Doctoral School of Electronics,

Telecommunications and Information Technology, National University for Science and Technology Politehnica Bucharest (NUSTPB). She graduated from the Faculty of Medical Engineering, NUSTPB and she holds a Master's Degree in ″Intelligent systems and computer vision″ from Faculty of Electronics, Telecommunications and Information Technology, NUSTPB. Her research fields and topics of interest include artificial intelligence applications, especially in the medical field (processing and analysis of medical images and medical signals), telemedicine applications and the development of healthcare equipment.

**Alina-Elena SULTANA** este conferențiar universitar la Facultatea de Electronică, Telecomunicații și Tehnologia Informației din cadrul Universității Naționale de Știință și Tehnologie Politehnica București. Are peste 17 ani de experiență în procesarea cercetării imagistice medicale și a vederii computerizate. A participat (ca investigator principal sau membru al echipei) la peste 15 contracte de cercetare cu finanțare națională (publică și privată) cât și europeană. Alina Sultana a fost coautor a două lucrări de jurnal ISI, coautor al unui brevet internațional și a peste 40 de publicații ale conferinței ISI: publons.com/researcher. Profilul ei de cercetare Google poate fi găsit la: https://scholar.google.com.sg/ citations?hl=en&user=5XgMVtMAAAAJ.

**Alina-Elena SULTANA** is an associate professor at the Faculty of Electronics, Telecommunications and Information Technology of the National University of Science and Technology Politehnica Bucharest. She has over 17 years of experience in processing research medical imaging and computer vision. She participated (as principal investigator or team member) to more than 15 research contracts with national (public and private) and European funding. She co-authored two ISI journal papers, co-author of an international patent and over 40 ISI conference publications: publons.com/researcher. Her research profile Google can be found at: https://scholar.google.com.sg/citations?hl=en&user=5XgMVtMAAAAJ.

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.