RESEARCH-ARTICLE

# Application of vision transformers in the early detection of excavation in the BRSET base

**JOEL SANTOS FERREIRA**, University of Trás-os-Montes and Alto Douro, Vila Real, Vila Real, Portugal

**MIGUEL M FERNANDES**, University of Trás-os-Montes and Alto Douro, Vila Real, Vila Real, Portugal

**DANILO D.L. LEITE**, University of Trás-os-Montes and Alto Douro, Vila Real, Vila Real, Portugal

**DIBET GONZALEZ**

**JOSE CARLOS J.C. RAPOSO DA CAMARA**, University of Trás-os-Montes and Alto Douro, Vila Real, Vila Real, Portugal

**JOÃO J.R. RODRIGUES**, University of Algarve, Faro, Faro, Portugal

**View all**

**Open Access Support** provided by:

**University of Trás-os-Montes and Alto Douro**

**University of Algarve**



**Published:** 31 July 2025



DSAI 2024: 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion
*November 13 - 15, 2024*
*Abu Dhabi, United Arab Emirates*

# Application of vision transformers in the early detection of excavation in the BRSET base

Joel Santos Ferreira
University of Trás-os-Montes and Alto Douro (UTAD)
Portugal
al75737@alunos.utad.pt

Miguel M Fernandes
University of Trás-os-Montes and Alto Douro (UTAD)
Portugal
al76635@alunos.utad.pt

Danilo D.L. Leite
University of Trás-os-Montes and Alto Douro (UTAD)
Portugal
danilol@utad.pt

Dibet Gonzalez
Sentinel-Concept
Portugal
dibetg@gmail.com

Jose Carlos J.C. Raposo da Camara
University of Trás-os-Montes and Alto Douro (UTAD)
Portugal
jcrcamara@hotmail.com

João J.R. Rodrigues
NOVA LINCS & ISE, University of Algarve
Portugal
jrodrig@ualg.pt

António A.C. Cunha
University of Trás-os-Montes and Alto Douro (UTAD)
Portugal
University do Algarve, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
Portugal
acunha@utad.pt

## Abstract

Enlarged excavation of the optic papilla, caused by the loss of fibres that originate in the retina and transmit electrical stimuli to the visual cortex, is a critical indicator in the early detection of glaucoma, a disease that can lead to irreversible blindness. As the optic papilla shows morphological variations in the population, its identification can be a challenge. Methods based on deep learning have shown promise in helping doctors analyse these images more accurately. Recently, models such as Vision Transformers (ViT) have shown significant results in various medical applications, including glaucoma detection. However, the scarcity of quality data remains a major obstacle to training these models. This study evaluated the performance of the Swin Transformer, DeiT and Linformer models in detecting optic papilla excavation, using the new Brazilian Multilabel Ophthalmological Dataset (BRSET). The results showed that the DeiT model obtained the best accuracy, with 0.94, followed by the Swin Transformer, with 0.88, and the Linformer, with 0.85. The findings of this study suggest that ViT models can not only significantly improve the detection of glaucomatous papillary excavation, but also strengthen Human-Machine Collaboration, promoting more effective interaction between doctors and automated systems in medical diagnosis.

## CCS Concepts

• **Computing methodologies**; • **Machine Learning**;

## Keywords

Deep learning, Brazilian Multilabel Ophthalmological Dataset, Image Classification, Opthalmology

**ACM Reference Format:**
Joel Santos Ferreira, Miguel M Fernandes, Danilo D.L. Leite, Dibet Gonzalez, Jose Carlos J.C. Raposo da Camara, João J.R. Rodrigues, and António A.C. Cunha. 2024. Application of vision transformers in the early detection of excavation in the BRSET base. In *11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2024), November 13–15, 2024, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696593.3696633

## 1 Introduction

Glaucoma is the leading cause of irreversible blindness in the world, affecting millions of people and profoundly impacting their quality of life. Most cases progress asymptomatically. The World Health Organisation estimates that by 2040, around 112 million people could be affected by this condition. Diagnostic screening for glaucoma involves several stages: ophthalmological examination, measurement of eye pressure, visual field tests and optical coherence tomography. Other tests may be requested. Media transparency of the eyeball



*DSAI 2024, November 13–15, 2024, Abu Dhabi, United Arab Emirates*

ACM ISBN 979-8-4007-0729-2/24/11
https://doi.org/10.1145/3696593.3696633

makes it possible to assess the papilla directly or using colour photographs. It is one of the most important steps in differentiating normal and glaucomatous papillae. Some characteristics stand out in increased excavation: areas of sectoral retinal atrophy, thinning of the macula, and presence of haemorrhages in the papilla [1, 2].

Specialised cells in the retina transform light into electrical stimuli transmitted via fibres that cross the retina and converge in the papilla, forming the optic nerve and heading towards the visual cortex where the electrical stimuli are interpreted. Increased excavation represents a diffuse or localised loss of these fibres and takes on different morphologies in the optic papilla. However, the presumptive diagnosis of glaucomatous disease is made when the loss of fibres corresponds to a loss of visual field. Diagnosis and follow-up treatment depend on the interpretation of ophthalmological data by the specialist, but the growing demand for ophthalmological examinations, combined with the shortage of specialists, highlights the need for automated methods for screening and early detection of various eye diseases.

The growing demand for ophthalmological examinations, combined with the shortage of specialists, highlights the need for automated methods for early detection of the disease to prevent progression to blindness and mitigate its socio-economic impact [3–5]. Human-machine collaboration (HMC) requires that machines in the broadest sense be designed to work together or learn how to work with humans. This means that hardware, software, and interfaces must be able to assess the user's unique needs and behaviours as well as the context in which they are used and the surrounding environment in order to enable on-the-spot cooperation with humans. This paper presents a method to detect glaucomatous disease, giving one more step for HMC.

In this scenario, the development of solutions based on deep learning (DL) models has significantly progressed and shown promising results. Among these advances, technologies such as transformers have stood out for improving diagnostic efficiency and providing more accurate and reliable results [6]. Vision Transformers (ViT) have shown excellent performance in analysing medical images, enabling fast and informed decisions due to their ability to capture global image characteristics through the self-attention mechanism [7, 8]. In this context, recent studies have explored the effectiveness of Transformers in detecting glaucoma. Fan et al.'s study [9] investigated the applicability of the Vision Transformer (DeiT) in detecting glaucoma using fundus images.

The results showed that DeiT achieved good levels of accuracy, outperforming ResNet-50 on both the OHTS test datasets and five external datasets: REFUGE, RIM-ONE, DRISHTI-GS and SCEH. Medvedeva and Kholod [10] explored the use of ViTs to detect glaucoma, demonstrating that integrating these networks with medical images can considerably improve the accuracy of diagnosis and the effectiveness of screening in ophthalmology. They highlighted the ability of ViTs to capture subtle and vital details in fundus images, contributing to a more accurate diagnosis.

Similarly, Leite et al. [3] compared ViT with other approaches in detecting patients with increased optic nerve cupping using the Brazilian Multilabel Ophthalmological Dataset (BRSET). Their results indicated that ViT models showed promising performance, outperforming other techniques. Additionally, Gould, Yang and Clifton [11] evaluated the performance of the ResNet-50 model trained on BRSET to classify various eye diseases, including glaucoma, automatically. They demonstrated the effectiveness of ResNet-50 in classifying these conditions but emphasised that Transformer-based models can offer further improvements in terms of accuracy and explainability.

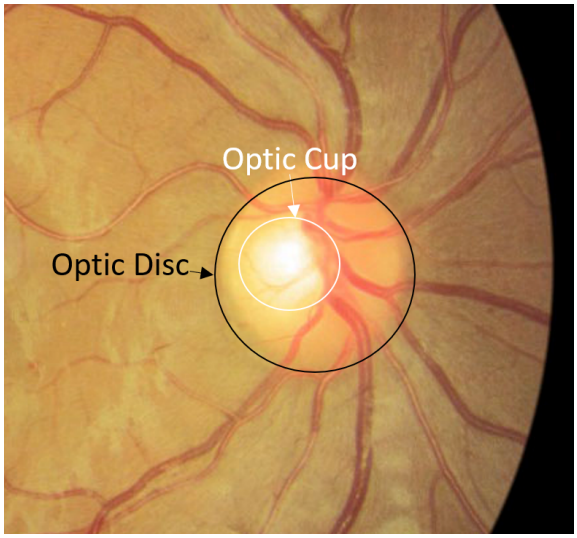


**Figure 1: - Eye fundus image with localised optic disc and cup (BRSet)**

On the other hand, the need for representative and secure datasets is fundamental for training and evaluating the effectiveness of these models [6]. In this scenario, the Brazilian BRSET dataset has recently become available, offering a vast amount of diverse, high-resolution images, facilitating more effective training and testing of the models, and allowing for better generalisation [9, 11]. Combining the features of BRSET with the capabilities of Vision Transformers can provide an ideal environment for developing accurate and generalisable models. Therefore, this study aims to evaluate and compare the results of three Vision Transformer (ViT) models in detecting and assessing optic papilla excavation using BRSET colour retinal images. The models tested include the Swin Transformer [12], DeiT [13] and Linformer [14] architectures.

Being the major contribution of the paper this comparison, for the best of our knowledge, is not yet done. In addition, these models were selected due to their superior performance in various computer vision tasks, ability to capture global image features, and efficiency in processing large volumes of visual data.

Next Section we present the materials and methods, this includes the datasets and data preparation and augmentation, Training and finally evolution metrics. Section 3 it is presented the results and discussion, and the final section presents the conclusions and future work.

## 2 Materials and methods

Fundus image with a localised optical disc and optical cup, illustrated in Figure 1. The workflow adopted in this study is detailed in Figure 2 and comprises three main stages: data collection and processing, training and model building, and performance evaluation.

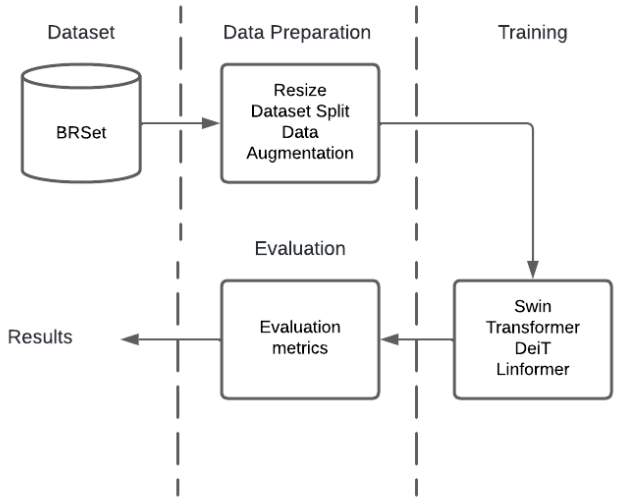


**Figure 2: - Pipeline for classification**

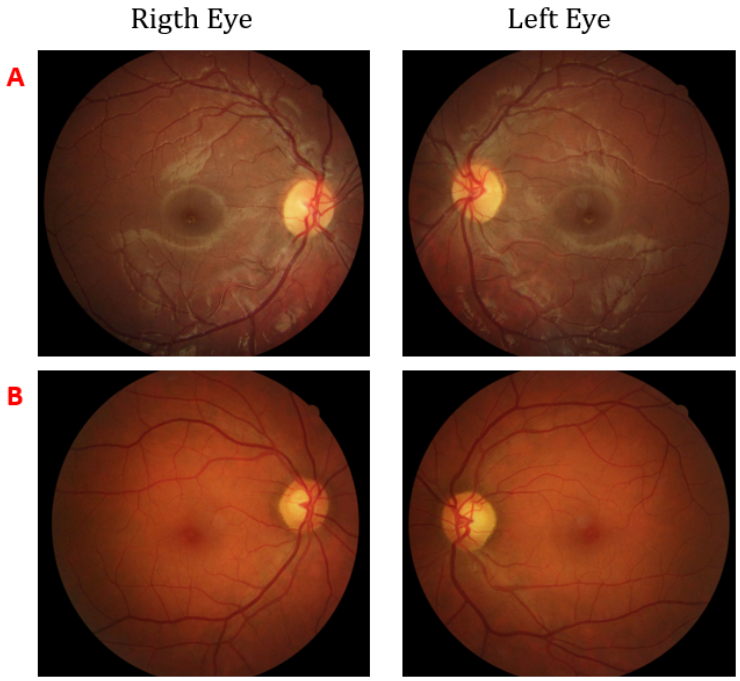


**Figure 3: - Sample images from the BRSet dataset of the background: (A) Absent, (B) Present**

This systematised method enables rigorous and objective analysis, crucial for developing more accurate diagnoses in ophthalmology. The block diagrams presented in Figure 2 will be detailed in the next subsections.

## 2.1 Datasets

In this study, we used the Brazilian Multi-Label Ophthalmological Dataset (BRSET), available on PhysioNet, made up of 16.266 retinal images from 8.524 Brazilian patients. [15]. This dataset includes images centred on the fovea, showing both visible temporal retinal vascular arches and at least one visible retinal disc diameter on the nasal side of the optic disc, as illustrated in Figure 3.

**Table 1: - List of the publicly available databases.**

| Dataset | Images | Resolutions | Source |
|---|---|---|---|
| BRSET | 16266 | 951x874 | [15] |
| ACRIMA | 705 | 2048x1536 | [17] |
| REFUGE | 1200 | 2124x2056, 1634x1634 | [18] |
| RIM-ONE | 455 | 2144x1424 | [19] |
| Drishti-GS1 | 101 | 2896x1944 | [20] |
| ORIGA-light | 650 | 3072x2048 | [21] |
| sjchoi86-HRF | 401 | 2592x1728 | [22] |

Of the available images, 3.202 were classified as positive for optic excavation and 13.064 as unfavourable, with an average patient age of 57.09 ± 18.1 years. BRSET is a valuable resource for eye disease research, allowing machine learning models to predict demographic characteristics and classify multilabel diseases using retinal images. In addition, the dataset includes critical clinical parameters such as enlarged optic nerve excavations, an essential factor in diagnosing glaucoma. This dataset is designed to enhance the development of the scientific community and validate machine learning models. We have included a comparative table based on information from different articles to provide a comparative perspective and highlight the characteristics of BRSET compared to other datasets available for glaucoma detection [16].

## 2.2 Data Preparation and Augmentation

Data preparation and augmentation are essential steps in developing DL models, ensuring that the data is in the correct format and that the model is trained with a representative data set [17, 18].

*2.2.1 Data Preparation.* All fundus images were prepared by removing file IDs, sensitive information, and headers and focusing on the macula for a consistent viewpoint. The images were resized to 224 x 224 pixels, according to Zhang [19], to balance quality and computational efficiency. The pixels were then normalised to mean 0 and standard deviation 1, standardising the scale of the data and avoiding bias, as suggested by Krizhevsky [20]. The colour channels were standardised for uniformity, according to Howard [21], and a Gaussian filter was applied to reduce noise, improving the quality and accuracy of the model, as recommended by Simonyan and Zisserman [22].

*2.2.2 Data Augmentation.* In medical image analysis, where data variability is high, and the amount of labelled data can be limited, traditional data augmentation techniques often don't fully capture the complexity of datasets. To address this challenge, we use Adaptive Data Augmentation (ADA), which dynamically adjusts augmentation strategies based on the specific characteristics of the dataset, providing more contextually relevant transformations. The transformations applied include horizontal inversions, rotation, scale adjustment, filling, cropping, adding Gaussian noise and perspective transformation. Lighting and colour parameters were adjusted, along with variations in sharpness and blur. ADA dynamically adapts these transformations to the specific characteristics of the data, changing, for example, the probability of inversions, rotation angles and scale factors according to the properties of the images. Specific adaptive transformations, such as Elastic Transform and Grid Distortion, are used to provide more relevant variations. ADA improves the effectiveness of data augmentation by customising transformations according to the intrinsic variability of the data, providing better generalisation and reducing the risk of overfitting [23].

## 2.3 Training

In the training process, the data set was randomly divided into three subsets: 70 per cent for training, 15 per cent for validation and 15 per cent for testing. This division is essential to ensure that the model is trained and evaluated in a representative manner. In addition, stratified cross-validation with six folds was applied to provide a robust and unbiased model evaluation. This method distributes the samples evenly in each fold, preserving the proportion of classes and minimising data imbalance, which avoids bias in assessing the model's performance.

For hyperparameter optimisation, the Optuna library played a pivotal role. This library efficiently and effectively adjusts the parameters, significantly contributing to the model's optimisation. Optuna was set to suggest values for the model's hyperparameters, including learning rate (1e-4), weight decay (between 1e-4 and 1e-2), beta1 and beta2 (between 0.8 and 0.999), as highlighted by Kahloot and Ekler [24]. The AdamW optimiser was used with the settings suggested by Optuna [25], along with a ReduceLROnPlateau scheduler to dynamically adjust the learning rate during training. The loss function used was BCEWithLogitsLoss, which is appropriate for binary classification. Training was carried out in 10 epochs for each fold in the cross-validation. This approach not only optimised the model's performance but also facilitated efficient learning and promoted good generalisation capacity on new data.

*2.3.1 Models. The models were selected because of their performance in various computer vision tasks, ability to capture global image characteristics, and efficiency in processing large volumes of visual data. They are:* Swin Transformer: is a neural network architecture for computer vision tasks. It uses a shifted windows approach, which improves computational efficiency by limiting the self-attention calculation to non-overlapping local windows. This technique allows the model to capture broader contextual information by allowing some interaction between windows while maintaining efficiency. Its hierarchical architecture divides the input image into small patches processed locally before combining this information at higher levels of the network. This method allows the model to adapt to different resolution scales, making it

ideal for image recognition and object detection tasks. In addition, the Swin Transformer is resource-efficient, making it suitable for systems with computational limitations. The Swin Transformer has demonstrated high accuracy in the detection of glaucoma, capturing detailed features of retinal images, which is crucial for early and accurate diagnosis [12, 18].

DeiT (Data-efficient Image Transformers): DeiT [13] has shown significant results in computer vision tasks and is promising in glaucoma detection. Unlike other models that require large amounts of data and substantial computational resources, DeiT offers good performance with less training data. DeiT's architecture divides images into small patches that are transformed into embeddings and inserted into a Transformer Encoder, which processes the relationships between the patches to understand the global structure of the image. A feature of DeiT is the introduction of two unique tokens: the Class Token, which represents the entire image and is used to make the final prediction, and the Distillation Token, which facilitates learning efficiency by allowing the model to learn from a previously trained model, such as ResNet. This approach improves training efficiency and model performance with less data [14]. Although designed to be efficient with less data, DeiT is highly scalable and can be used with large amounts of data while maintaining its effectiveness and accuracy. The transformation of patches into embeddings allows the model to process large data sets in a parallel and efficient way. It takes advantage of access to a larger volume of data to further improve its performance and generalisation. In the context of glaucoma detection, DeiT's efficiency in using data in an optimised way is especially relevant. Glaucoma is an eye disease diagnosed by analysing retina images and looking for subtle signs such as optic nerve excavation. DeiT's ability to capture a wide variety of features and patterns in retinal images is crucial for the accurate detection of these conditions.

Linformer: It is an architecture designed to improve the efficiency of attention models by reducing the computational complexity traditionally associated with the self-attention mechanism. The technique proposed by Wang [14] linearises the complexity of attention, making processing more feasible for large sequences and data sets. Traditionally, self-attention has a quadratic complexity about the length of the input sequence, which can be unfeasible for large sequences. Linformer solves this problem by applying a linear projection to the keys and values, reducing the dimensionality and resulting in a lower-dimensional approximate attention matrix. This reduced attention matrix is then used to calculate the dependencies between the input sequence elements while maintaining the ability to capture global relationships. By reducing the dimensionality of attention matrices, Linformer significantly decreases memory usage and computing time, making it suitable for large data sets and long sequences and significantly reduces memory usage and computing time without compromising the model's accuracy, making it ideal for applications with limited resources and large volumes of data. Recent studies highlight that Linformer can capture global dependencies in the input sequence, making it effective in natural language processing and computer vision tasks. With these improvements, Linformer is positioned as an efficient solution to the challenges of attention modelling in data-intensive scenarios. Even with the reduction in dimensionality, it retains the ability to capture global dependencies in the input sequence, which is essential for tasks such as analysing retinal images.

## 2.4 Evolution Metrics

Accuracy is fundamental in evaluating machine learning models, covering two essential dimensions: discrimination and reliability. Discrimination refers to the model's ability to correctly differentiate between categories of data, while reliability concerns the consistency of the predictions made over time. To assess these dimensions, we use a set of specific metrics, which are particularly effective for understanding model performance in detail. In this study, we chose to employ the following metrics: Accuracy (1), Precision (2), Recall (3), F1 Score (4).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$F1 = 2 * \frac{\mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

Each of these metrics is defined by corresponding equations, where TP (True Positives), TN (True Negatives), FP (False Positives) and FN (False Negatives) categorise the samples as "True Positives", "True Negatives", "False Positives" and "False Negatives", respectively. These categories are essential for cases in which the model (i) correctly identifies pixels or regions of interest (True Positives), (ii) correctly determines non-relevant pixels or regions (True Negatives), (iii) wrongly classifies non-relevant pixels as relevant (False Positives) and (iv) fails to identify pixels or regions that are relevant (False Negatives). In this phase, we used the following classification to evaluate the Precision, Recall, F1-score Accuracy (> 0.90), good (0.80 - 0.90), acceptable (0.70 - 0.80), poor (0.60 - 0.70) and no acceptable discrimination capacity (< 0.60) metrics.

# 3 Results and Discussion

The effectiveness of the models in accurately and sensitively detecting optic disc cupping was assessed using performance metrics such as accuracy, F1 score, precision and recall, categorised on scales ranging from excellent to unacceptable. A score above 0.90 was considered exceptional, reflecting the models' high discrimination capacity. Performances between 0.80 and 0.90 were promising, indicating adequate effectiveness in identifying excavation cases. Results between 0.70 and 0.80 were assessed as acceptable, suggesting adjustments needed to improve sensitivity and specificity. Results between 0.60 and 0.70 were considered poor, demonstrating that performance needs significant interventions to be effectively helpful in clinical practice. Any result below 0.60 was classed as unacceptable, demonstrating the models' inability to discriminate between excavation cases adequately. The results obtained in this study highlight the models' ability to distinguish cases with indications of excavation, as detailed in Table 2.

The results of this study indicate that the Swin Transformer, DeiT, and Linformer models effectively identify cases of optic disc cupping, a critical factor in screening for eye diseases that can lead to vision loss. Despite the unbalanced dataset, i.e., the ratio

**Table 2: – Results.**

| Models | Class | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| Swin Transformer | 0 | 0.93 | 0.91 | 0.91 | 0.88 |
| | 1 | 0.74 | 0.73 | 0.73 | |
| DeiT | 0 | 0.94 | 0.98 | 0.96 | 0.94 |
| | 1 | 0.92 | 0.75 | 0.82 | |
| Linformer | 0 | 0.88 | 0.90 | 0.91 | 0.85 |
| | 1 | 0.71 | 0.73 | 0.70 | |

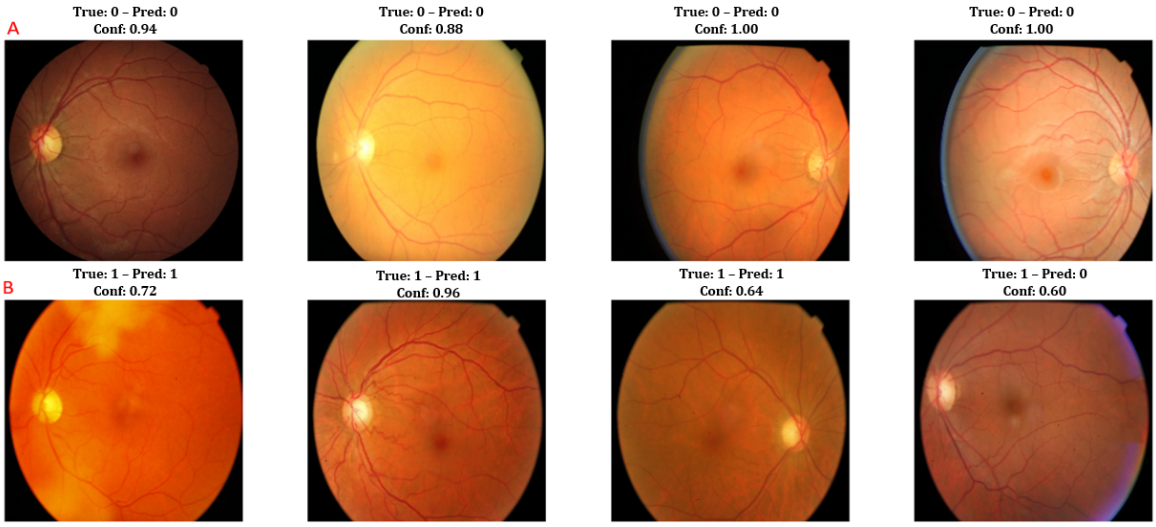


**Figure 4: - Model predictions in identifying optical disc excavation**

between papillae without cupping and those with increased cupping is around 25 per cent. An unbalanced dataset can significantly affect the performance of machine learning models with a low ability to correctly identify positive (or sick) cases, resulting in a higher number of false negatives.

DeiT stood out with the best results. Class 1 (presence of excavation) obtained a precision of 0.92, recall of 0.75 and F1 score of 0.82. For Class 0 (absence of excavation), the values were 0.94, 0.98 and 0.96, respectively, resulting in an overall precision of 0.94. Despite the unbalanced data set, with approximately 25 per cent of the papillae with the greatest excavation, DeiT managed to maintain a high performance. This is crucial, as unbalanced datasets generally lead to a higher number of false negatives. DeiT's ability to capture various features in retinal images is critical to detecting these conditions accurately. Figure 4 of the study shows a sample of the model's predictions, organised into two sections: A and B. Section A's cases belong to Class 0 (absence of excavation). The model correctly identified the lack of excavation in all the cases presented, with high confidence levels ranging from 0.88 to 1.00.

This performance can be attributed to DeiT's ability to capture subtle details and patterns in retinal images using the knowledge distillation approach and the auto-attention mechanism. These features allow the model to effectively understand the relationships between different parts of the image, resulting in accurate and reliable predictions. Section B's cases belong to Class 1 (presence of excavation). Although the model could correctly identify most excavation cases, the confidence levels ranged from 0.60 to 0.96. In one case, the model failed to predict the absence of excavation, even though excavation was present, with a confidence level of 0.60. This may identify a difficulty for the model in dealing with the minority class, which is the presence of excavation. This difficulty can be attributed to the imbalance of the data set and the intrinsic complexity of detecting subtle signs of excavation. However, these results show the efficiency of DeiT in detecting the absence of excavation.

In Class 0, the Swin Transformer showed a precision of 0.93, a recovery of 0.91 and an F1 score of 0.91, while in Class 1, the values were 0.74, 0.73 and 0.73, respectively. Despite its good performance, Swin Transformer showed lower precision than DeiT when applied to unbalanced datasets. However, it is still a solid option for tasks that require a detailed understanding of local structures in medical images [26, 27]. Linformer, designed to be memory and computationally efficient by reducing the complexity of auto attention linearly concerning sequence size, performed well overall, with a precision of 0.88, a recovery of 0.90 and an F1 score of 0.91 in Class 0 and 0.71, 0.73 and 0.70 in Class 1, resulting in an overall precision of 0.85. Linformer's main advantage is its scalability and

efficiency, making it suitable for applications with limited computing resources. Transformers, in general, have demonstrated significant advantages over traditional convolutional neural networks (CNNs) in various computer vision tasks. Transformers' ability to capture long-range dependencies and flexibility to scale with large amounts of data contribute to their superior performance in medical imaging applications. Therefore, the results of this study suggest that DeiT is a promising tool for the screening and early diagnosis of optic disc-related conditions. DeiT's high precision and generalisability guarantee timely and accurate diagnoses, optimising the treatment and management of patients with glaucoma or other eye diseases [28].

## 4 Conclusion

The results of this study demonstrate the effectiveness of ViT models in detecting excavation of the optic papilla, which is essential for diagnosing glaucoma. The Swin Transformer, DeiT and Linformer architectures performed well in discriminating between positive and negative images for optic excavation, even with a highly unbalanced data set. Among the models tested, DeiT stood out, with a precision of 0.94, recall of 0.98 and F1-score of 0.96 for Class 0 (absence of excavation) and a precision of 0.92, recall of 0.75 and F1-score of 0.82 for Class 1 (presence of excavation), reinforcing its potential as an effective screening tool in the ophthalmological context. The Swin Transformer and Linformer models also showed good results in screening the optic papilla. Implementing these models in clinical settings could significantly improve the efficiency of ophthalmological diagnoses, providing more accurate diagnoses and timely interventions.

Future work includes the continuous development and improvement of ViT architectures, with a view to expanding their applications beyond glaucoma and improving the clinical management of patients with other conditions related to optic disc health. In addition, an important area for research is the combination of ViTs with other deep learning techniques, such as convolutional neural networks (CNNs), to develop hybrid architectures that offer a balance between accuracy and computational efficiency. These initiatives not only have the capacity to optimise the clinical management of glaucoma patients, but also to extend the use of ViT-based technologies to a wider variety of eye diseases.

## Acknowledgments

This work is funded by National Funds through the Portuguese Foundation for Science and Technology (FCT), and supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020) and ref. UIDP/04516/2020 (https://doi.org/10.54499/UIDP/04516/2020) with financial support from FCT.IP and INESC - This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

## References

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, e C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis", Ophthalmology, vol. 121, no 11, p. 2081–2090, 2014.

[2] J. Camara, R. Rezende, I. M. Pires, e A. Cunha, "Retinal Glaucoma Public Datasets: What Do We Have and What Is Missing?", J. Clin. Med., vol. 11, no 13, p. 3850, 2022.

[3] D. Leite, J. Camara, J. Rodrigues, and A. Cunha, "A Vision Transformer Approach to Fundus Image Classification."

[4] J. Camara, A. Neto, I. M. Pires, M. V. Villasana, E. Zdravevski, e A. Cunha, "Literature Review on Artificial Intelligence Methods for Glaucoma Screening, Segmentation, and Classification", J. Imaging, vol. 8, no 2, p. 19, 2022.

[5] L. M. Zangwill, M. D. Bowd, C. Girkin, N. Weinreb, and R. A. Medeiros, "Optic nerve head and retinal nerve fiber layer analysis: A review," Journal of Glaucoma, vol. 16, no. 3, pp. 288-298, May 2007.

[6] A. Elmoufidi and S. Jai-Andaloussi, "Transformers for Medical Image Analysis - Applications, Challenges, and Future Scope," ResearchGate, 2023

[7] H. Zhang and Y. Qie, "Applying Deep Learning to Medical Imaging: A Review," Applied Sciences, vol. 13, no. 18, p. 10521, 2024.

[8] D. Leite *et al.*, "Machine Learning automatic assessment for glaucoma and myopia based on Corvis ST data," in Procedia Computer Science, Elsevier B.V., 2021, pp. 454–460.

[9] R. Fan *et al.*, "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," Ophthalmol. Sci., vol. 3, no. 1, p. 100233, 2023.

[10] E. Medvedeva and S. Kholod, "Vision Transformers in Human Vision Analysis Including Glaucoma Detection," IEEE Transactions on Medical Imaging, vol. 43, no. 5, pp. 1256-1267.

[11] D. S. W. Gould, J. Yang, and D. A. Clifton, "Deep Learning for Multi-Label Disease Classification of Retinal Images: Insights from Brazilian Data for AI Development in Lower-Middle Income Countries," medRxiv, 2024.

[12] P. Y. Kim, J. S. Kwon, S. Joo, S. P. Bae, D. Lee, S. Yoo, J. Cha, and T. Moon, "SwiFT: Swin 4D fMRI Transformer," 2023.

[13] "DiT-3D: Exploring Plain Diffusion Transformers for 3D Shape Generation," 2023.

[14] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-Attention with Linear Complexity," arXiv.org, 2020.

[15] L. F. Nakayama *et al.*, "A Brazilian Multilabel Ophthalmological Dataset (BRSET) v1.0.0," 2023. https://physionet.org/content/brazilian-ophthalmological/1.0.0/.

[16] L. Deininger, B. Stimpel, A. Yüce, S. Abbasi-Sureshjani, S. Schönenberger, P. S. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and CNNs in digital pathology," arXiv.org, 2022.

[17] Elmoufidi, Abdelali, and Said Jai-Andaloussi. "CNN with Multiple Inputs for Automatic Glaucoma Assessment Using Fundus Images." International Journal of Image and Graphics, 2022.

[18] J. Hernández, F. J. Fumero, J. F. Sigut, and T. Díaz-Alemán, "Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images," Appl. Sci., vol. 13, no. 23, p. 12722, 2023.

[19] Zhang *et al.*, "Discussion on image resizing techniques in deep learning models," Journal of Image Processing, 2020.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Introduction to pixel normalization to improve convergence in deep neural networks," in Neural Information Processing Systems, 2012.

[21] J. Howard *et al.*, "The importance of color channel standardization in deep learning models," in International Conference on Learning Representations, 2017.

[22] K. Simonyan and A. Zisserman, "Application of Gaussian filters for noise reduction in images," International Journal of Computer Vision, 2015.

[23] Kim, M., and Bae, H.-J., "Data Augmentation Techniques for Deep Learning-Based Medical Image Analyses," Journal of the Korean Society of Radiology, vol. 81, no. 6, pp. 1290–1304, 2020, doi: 10.3348/jksr.2020.0158.

[24] S. T. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," Sinkron: jurnal dan penelitian teknik informatika, vol. 7, no. 4, pp. 11792, 2022.

[25] Kahloot and Ekler, "Hyperparameter optimization using Optuna," 2021.

[26] Y. Ma, Q. Yan, Y. Liu, J. Liu, J. Zhang, and Y. Zhao, "StruNet: Perceptual and low-rank regularized transformer for medical image denoising," Medical Physics, 2023. doi: 10.1002/mp.16550.

[27] J. You, M. K. Hasan, and Y. J. Zhang, "Transformers in medical image analysis," Medical Image Analysis, vol. 76, p. 102294, 2022.

[28] Y. Chen *et al.*, "Transforming medical imaging with Transformers? A comparative review," arXiv preprint arXiv:2106.01072, 2021.