Instability and Interpretability Discrepancies Between CNNs and Vision Transformers in Keratoconus Detection

Anonymous Authors[1]

Abstract

The stability and reliability of explanations from explainable artificial intelligence (XAI) are unclear, while its use has grown in medical imaging. In this study, we evaluate the stability and faithfulness of class activation mapping (CAM) based explanations from convolutional neural networks (CNNs) and vision transformers (ViTs) when classifying keratoconus with corneal topography maps. CNNs (ResNet-50) and ViTs (B/16) were trained on a public dataset of 4,011 corneal topography images classified as normal, suspect, or keratoconus, with all experiments being conducted across 7 random seeds to improve reproducibility. Grad-CAM generated heatmaps displaying regions important to a model's diagnosis. Explanation reliability was measured with CAM instability, structural similarity index (SSIM), performance with CAM and random pixel removal, and explanation-driven vulnerability (EDV). ViTs showed higher cross-seed explanation similarity than CNNs (SSIM: 0.604 vs 0.456). The ViT had an order-of-magnitude reduction in instability compared to the CNN (0.0001 vs. 0.0013). ViT also had a $4\times$ reduction in EDV compared to the CNN (0.0003 vs. 0.0012), while CNN prediction confidence was more sensitive to the removal of CAM-guided regions compared to random regions initially (top-k 10% to top-k 30% removal: 0.568 to 0.349 (CNN) vs. 0.863 to 0.861(ViT)), suggesting CNN explanations are more faithful but more variable, while ViT explanations are less faithful but more stable. These findings suggest that CAM reliability is strongly architecture-dependent, and that XAI should be evaluated before clinical use.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

1.1. Keratoconus Prevalence and Significance

Keratoconus is a progressive eye disease characterized by the gradual thinning and bulging of the cornea, resulting in irregular astigmatism and visual distortion (Cleveland Clinic, 2023). Some of the largest risk factors for the disease include eczema, allergy, and a family history of keratoconus. The prevalence of keratoconus also varies significantly across populations, with reports suggesting a prevalence of 0.2 to 4,790 per 100,000 persons, while its incidence lies between 1.5 and 25 cases per 100,000 persons/year. It has been identified as common in 20 to 30-year-old individuals with Middle Eastern and Asian descent (Santodomingo-Rubido, Carracedo, Suzaki, Villa-Collar, Vincent, and Wolffsohn, 2022).

1.2. An Overview of Corneal Topography

Corneal topography is an imaging technique using topography scans to map a 3D layout of the cornea. Using computer-assisted mapping and by evaluating the curvature, elevation, and thickness of the cornea, this technology can be used for injuries scarring the cornea, astigmatism, contact lenses, unwanted growths, and keratoconus (Porter, 2021).

Current parameters to study keratoconus using corneal topography include mean keratometry (K), astigmatism, and corneal elevation. However, these parameters have often failed in suspected or subclinical keratoconus cases; for example, mean keratometry was reported to have no ability to significantly differentiate outcomes between keratonic and healthy eyes. Furthermore, the specificity of astigmatism as a diagnostic parameter was reported to decrease to less than 65% with subclinical keratoconus, also indicating that it is not an effective diagnostic parameter (Martínez-Abad and Piñero, 2017). Finally, Jafarinasab et al. find that the area under the curve (AUC) of clinical keratoconus (KCN) was 0.97, compared to an AUC for subclinical KCN of 0.69. However, for these metrics, the sensitivity was 89.23% for KCN, while it was only 50% for the subclinical KCN (Jafarinasab, Shirzadeh, Feizi, Karimian, Akaberi, and Hasanpour, 2015). These results, along with the earlier limitations, demonstrate the limitations of detecting early-stage

keratoconus with just single threshold-based topographic parameters. Thus, there may be a higher risk for a missed early diagnosis or increased risk of disease prognosis due to corneal topography's low sensitivity rate in subclinical keratoconus.

1.3. Machine Learning in Ophthalmology

In recent years, the use of artificial intelligence (AI) has only grown in popularity throughout the medical field. The technology has been utilized for image analysis, particularly in oncology, where it has enabled the detection of malignant melanoma from skin images. In ophthalmology, deep learning (DL) within AI can be used for the diagnosis of ocular diseases and telecare in clinical settings (Ting, Pasquale, Peng, Campbell, Lee, Raman, Tan, Schmetterer, Keane, and Wong, 2019).

Mushin et al. find that only 16% of ML studies of keratoconus utilize Convolutional Neural Networks (CNNs), with only 7% of studies analyzing both subclinical keratoconus and clinical keratoconus (Muhsin, Qahwaji, Ghafir, AlShawabkeh, Al Bdour, AlRyalat, and Al-Taee, 2025).

While many ML models report high diagnostic accuracy, Muhsin et al. emphasize that few approaches have been successfully translated into real-world clinical practice. One of the largest reasons for this is that ML models can have limited interpretability, since the review finds that most studies do not investigate how models analyze regions of the cornea in determining a final diagnosis; rather, they primarily focus on the end diagnosis (Muhsin et al., 2025).

Thus, studies must also evaluate ML in terms of interpretability, stability, and sensitivity, in combination with overall performance, to determine their true effectiveness.

1.4. Differences Between the Effectiveness of CNNs and ViTs

While CNNs are common in the medical field due to their architecture enabling them to extract local features, Vision Transformers (ViTs) are becoming more prominent in ophthalmology due to their performance at specific tasks exceeding that of CNNs (Takahashi, Sakaguchi, Kouno, Takasawa, Ishizu, Akagi, Aoyama, Teraya, Bolatkan, Shinkai, Machino, Kobayashi, Asada, Komatsu, Kaneko, Sugiyama, and Hamamoto, 2024).

The best ViT model created by Dosovitskiy et al. had an accuracy rate of 90.72% on ImageNet-ReaL, and an accuracy of 94.55% on CIFAR-100. These high accuracy rates demonstrate that pure transformers that are pretrained on large datasets can have performance similar to the state-of-the-art CNN models, establishing transformers, and specifically, Vision Transformers, as strong alternatives to CNNs (Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit, and Houlsby, 2021).

Vision Transformers are also better at capturing global contextual information due to self-attention and have more uniform representations across all of their layers. However, CNNs rely greater on pooling operations and stacked convolutional filters in order to extract local features that can gradually build global representations (Raghu, Unterthiner, Kornblith, Zhang, and Dosovitskiy, 2022). Understanding the difference in architecture between CNNs and Vision Transformers (ViTs) is necessary for clinical use, as these different models vary significantly in spatial feature capturing, thus impacting diagnoses.

Explainable artificial intelligence (XAI) is a type of machine learning model. It offers explanations that humans can understand for the predictions made by deep learning models. These serve to address the "black box problem" in artificial intelligence, where humans can provide inputs and receive outputs from deep learning models, but do not understand how the model achieved that output due to complex neural network processing. The use of XAI can provide trust and accountability to machine learning models when a human reviewer can understand how a model came to a specific decision, and make changes to parameters based on those results (Nasim, Ferdous, Rashid, Soshi, Biswas, Biswas, and Gupta, 2025).

Salience-based visualization, which highlights regions of an image that contribute the strongest in determining a model's prediction, is one of the most widely used categories of XAI methods for deep learning. Among these, class activation maps (CAMs) and variants like Gradient-weighted Class Activation Mapping (Grad-CAM) are growing in popularity as they can generate spatial heatmaps indicating areas of high importance to the model's diagnosis in an image. This allows clinicians to determine whether a model's attention to a specific region corresponds to relevant anatomical structures, such as in keratoconus (Selvaraju, Cogswell, Das, Vedantam, Parikh, and Batra, 2017).

Despite their widespread adoption, studies have also raised concerns about the reliability of CAM-based explanations, such as salience-based visualization. Adebayo et al. find that saliency maps can appear to be visually similar, even when model parameters are randomized, thus suggesting that some of the model's explanations might not truly reflect its full decision-making process (Adebayo, Gilmer, Muelly, Goodfellow, Hardt, and Kim, 2018).

In ophthalmology, the reliability of XAI is particularly important, since diagnoses often rely on spatial patterns in corneal images. Misleading explanations could lead to incorrect diagnoses by clinicians, even when model performance appears to be acceptable.

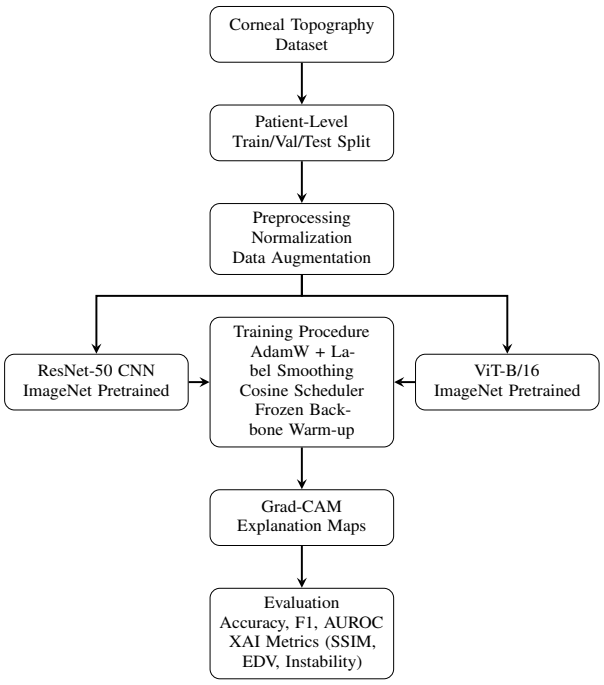


Figure 1. An outline of the methodology used in this study. After preprocessing, ImageNet pretrained weights are applied, with AdamW, a Cosine Scheduler and a Frozen Backbone Warm-Up being used during training. Metrics were calculated following the creation of Grad-CAM explanation maps.

These limitations necessitate assessments of explanation stability, perturbation sensitivity, and cross-architecture agreement. In particular, contrasting CNNs and ViTs provides insight into how the different architectures of these models influence attention mechanisms and the reliability of spatial explanations. By quantitatively analyzing CAM stability, cross-model similarity, and explanation-driven vulnerability, this study aims to assess the extent to which XAI methods can be trusted for corneal topography-based classification of keratoconus.

Thus, this study makes 3 main scientific contributions: (1) quantifying CAM instability across various keratoconus classes, (2) evaluating explanation agreement between CNNs and ViTs, and (3) assessing explanation faithfulness by using perturbation-based sensitivity metrics.

2. Methods

2.1. Dataset and Preprocessing

This study uses a dataset of corneal topography images, specifically containing normal, subclinical keratoconus, and clinical keratoconus cases. Every image uses posterior corneal elevation maps that are relative to a best-fit sphere reference, and all images were resized to a fixed resolution. They were also normalized to ensure consistent distributions of intensity across various samples. We also applied data augmentation techniques like horizontal flipping and slight rotations to reduce inter-subject variability during training.

The dataset used in the study was obtained from Kaggle (Hammouch). The dataset contains 4,011 images in total and is divided into three classes of keratoconus status: normal, suspect, and keratoconus. This dataset is public and de-identified, and was originally collected under approval of the IRB of the Federal University of São Paulo (UNIFESP/EPM) that follows the Declaration of Helsinki and its later amendments, and received informed consent from its participants.

The dataset was organized in case-level directories, where each directory had a unique subject and contained multiple images from the same patient. To prevent data leakage, we performed a group-wise split creating grouping variables by using case identifiers, thus ensuring that all images from a specific patient were confined to a specific partition (training, validation, or testing). To maintain a balanced evaluation, class distributions were preserved across splits.

2.2. Model Architectures

This study evaluates two deep learning architectures: a CNN (ResNet-50) and a ViT (B/16). We designed the CNN architecture to extract hierarchical spatial features by using successive convolutional and pooling layers, while we designed the ViT model to use a patch-based embedding scheme along with multi-head self-attention to analyze long-range dependencies in corneal topography images. To improve convergence and generalization, both models were initialized with pretrained weights from ImageNet and fine-tuned on the corneal topography dataset. We then used a fully connected classification head to produce class probabilities for each diagnostic category.

For the CNN, the ResNet-50 backbone used residual bottleneck blocks followed by global average pooling to extract hierarchical spatial features. We also replaced the original 1000-class fully connected layer with a dropout layer ($p = 0.4$) and a linear layer mapping the 2048-dimensional feature vector:

$$\hat{y} = W f(x) + b, \quad W \in \mathbb{R}^{3\times 2048}. \tag{1}$$

For the ViT, its B/16 model partitioned each image into 16×16 patches, which were embedded in a 768-dimensional space and then processed through 12 transformer encoder blocks. The final CLS token representation was passed through a dropout layer ($p = 0.4$), and a linear layer with three output logits:

$$\hat{y} = W z_{\text{CLS}} + b, \quad W \in \mathbb{R}^{3\times 768}. \tag{2}$$

2.3. Training Procedure

We trained the models using categorical cross-entropy loss with label smoothing ($\epsilon = 0.1$) and optimized them with the AdamW optimizer (CNN initial learning rate = 3×10^{-4}, ViT initial learning rate = 5×10^{-5}, weight decay = 1×10^{-4}). The following equation was used to calculate the categorical cross-entropy loss:

$$\tilde{y}_{n,i} = \begin{cases} 1 - \epsilon & \text{if } i = y_n \\ \frac{\epsilon}{C-1} & \text{otherwise} \end{cases} \tag{3}$$

The loss becomes:

$$L = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{C} \tilde{y}_{n,i} \log(\hat{y}_{n,i}) \tag{4}$$

where y_n represents the ground-truth class index for sample n, $\tilde{y}_{n,i}$ represents the smoothed target distribution, and $\hat{y}_{n,i}$ represents the softmax probability predicted.

We originally empirically selected learning rates, and used a cosine learning rate scheduler ($T_{max} = 60$). To prevent overfitting, early stopping with a patience of 8 epochs was used based on validation accuracy, with the model with the highest validation accuracy being saved for evaluation.

For the CNN, Grad-CAM was calculated with the final convolutional layer before global pooling, with gradients being obtained with respect to the predicted class. The feature maps produced were then linearly combined using gradient-based weights and used bilinear interpolation to resample to the input resolution.

For the ViT, Grad-CAM was implemented by registering gradient hooks on the final transformer block before the classification lead. The class token was also excluded from spatial attribution, and patch embeddings were reshaped into a 2D spatial grid before bilinear interpolation to input dimensions. All experiments were repeated across 7 random seeds, with a batch size of 16 being used for training/evaluation. For the first 5 epochs, the pretrained backbone remained frozen while the classification head was trained. After this, the pretrained backbone was unfrozen. The models were trained for a maximum of 60 epochs, and both architectures were initialized with ImageNet-pretrained weights. During training, data augmentation included random resized cropping (224x224), horizontal flipping, random rotation (±15°), and color jittering.

The models were evaluated using overall accuracy, balanced accuracy, macro-averaged F1 score, per-class F1 scores, confusion matrices, and multiple class AUROC. Explanations were evaluated with CAM instability under Gaussian perturbations, cross-architecture CAM instability (SSIM), top-k CAM-guided modification sensitivity, and Explanation-Driven Vulnerability (EDV).

2.4. Explainability and Quantitative Evaluation

We generated Grad-CAM explanations with respect to the predicted class c as:

$$\text{CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \quad \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

where A^k denotes the k-th feature map and Z is the spatial normalization factor.

Explanation stability under Gaussian perturbations $\delta \sim \mathcal{N}(0, \sigma^2)$ was quantified as:

$$\text{Instability} = 1 - \text{SSIM}(\text{CAM}(x), \text{CAM}(x + \delta)) \tag{6}$$

Gaussian noise $\delta \sim \mathcal{N}(0, 0.01^2)$ was added to each input image.

Under Explanation-Driven Vulnerability (EDV), confidence degradation was calculated with the following formulas:

$$\Delta_k = \left| p_c(x) - p_c(x^{(k)}) \right| \tag{7}$$

$$\text{EDV} = \text{Var}_{k \in \mathcal{K}}(\Delta_k) \tag{8}$$

with $x^{(k)}$ representing the input after masking the top-k most salient Grad-CAM regions.

We used Grad-CAM with both the CNN and ViT models to interpret the model predictions and to improve the consistency of results. For each image, Grad-CAM maps were generated with respect to the predicted class. Explanations were also generated at the final convolution layer for CNNs and at the last attention block for ViTs to ensure compatibility across different models. To assess explanation stability, we evaluated Grad-CAM consistency under controlled perturbations and applied small noise perturbations for each input image. The Structural Similarity Index Measure (SSIM) was also used to quantify the similarity between the original and perturbed explanation maps.

Furthermore, we compared the explanation maps generated by CNNs and ViTs with the same inputs to evaluate cross-architecture agreement. A low agreement between models was interpreted as a potential indicator of explanation unreliability. We performed explanation-guided occlusion experiments to determine if the explanations corresponded to clinically meaningful regions by ranking pixels on their normalized CAM intensity, with the top-k highest scoring pixels being set to 0 (hard deletion). We evaluated $k \in \{0.1, 0.2, 0.3, \ldots, 0.9\}$, which corresponds to masking 10-90% of the most salient pixels in 10% increments. For each image, EDV was defined as the variance of the deletion

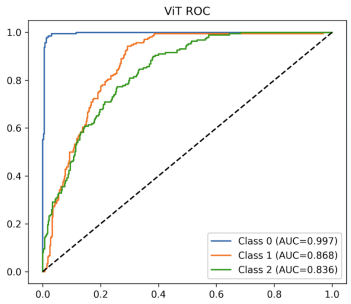
curve of the absolute drop in confidence across all k values. However, results of $k \in \{0.1, 0.3, 0.5, \ldots, 0.9\}$ are shown for simplicity. We interpreted a significant performance drop after occlusion as the possibility that the model relies on the highlighted regions for decision-making. Figure 1 represents the methodology pipeline used.

3. Results

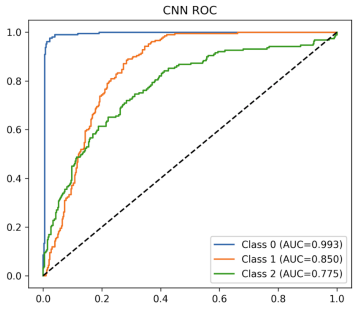
Table 1. Summary of multi-seed performance and explainability robustness metrics (mean ± SD across 7 seeds).

Metric	Value
Test Set Accuracy (CNN)	0.8384 ± 0.0580
Test Set Accuracy (ViT)	0.8299 ± 0.0431
Across-Seed SSIM (CNN)	0.4558 ± 0.1030
Across-Seed SSIM (ViT)	0.6039 ± 0.1129
EDV (CNN)	0.0012 ± 0.0019
EDV (ViT)	0.0003 ± 0.0006
Instability (CNN)	0.0013 ± 0.0004
Instability (ViT)	0.0001 ± 0.0001

3.1. Classification Performance



(a) ViT ROC



(b) CNN ROC

Figure 2. ROC curves for the ViT and CNN models.

The CNN had an overall classification accuracy of 0.8384 ± 0.0580, while the ViT had an overall classification accuracy of 0.8299 ± 0.0431 across the 7 random seeds tested (Table 1). Accuracy differences were not statistically significant (p = 0.813, Paired Cohen's d_z = 0.258), setting up a fair comparison for both models due to comparable performances (Table 3). Figure 2(a) and Figure 2(b) represent the ROC curves for the ViT and CNN models.

Due to similar performances, traditional performance metrics like classification accuracy are insufficient in fully evaluating a model's effectiveness at using corneal topography for keratoconus identification, motivating a deeper analysis using explainable artificial intelligence (XAI) as CNNs and ViTs become more commonly used in medical applications.

3.2. Explanation Instability

CAM instability was computed for the CNN and the ViT to quantitatively evaluate the robustness of the models' explanations. We measured instability as the variability of CAM outputs with small input perturbations, with higher values indicating less stable explanations. The reported CNN instability values were 0.0013 ± 0.0004, with ViT instability values reported as 0.0001 ± 0.0001 (p = 0.016, Paired Cohen's d_z = 2.531) (Table 1).

3.3. Cross-Seed CAM SSIM

This study used SSIM to compare the saliency maps produced by the respective models trained with random seeds (pairwise similarity was calculated between seeds). A higher SSIM value represents more stable, reproducible explanations. The CNN reported an SSIM of 0.4558 ± 0.0163, while the ViT reported an SSIM of 0.6039 ± 0.0241 (Table 3). Statistical analysis (p = 9.54×10^{-7}, W = 0, r = 0.88) found a highly statistically significant increase in explanation stability in the ViT across random seeds. Notably, all seed-pair comparisons favored the ViT, indicating a consistent increase in explanation stability compared to the CNN.

3.4. CNN Sensitivity Sweep (top-k CAM removal)

Using a progressive explanation sensitivity test, the model sensitivity of the CNN and ViT was measured while an increasing percentage of the image was removed (Table 2).

Table 2. CNN and ViT accuracy after top-k-guided CAM vs. random region removal (mean across 7 seeds).

	CNN		ViT	
Rem.(%)	**CAM**	**Rand.**	**CAM**	**Rand.**
10%	0.568	0.790	0.863	0.867
30%	0.349	0.643	0.861	0.838
50%	0.307	0.481	0.793	0.798
70%	0.304	0.325	0.623	0.730
90%	0.320	0.299	0.380	0.488

While both models reported a reduction in accuracy as in-

Table 3. Statistical analysis of multi-seed performance and explainability robustness, with p-values being calculated with Wilcoxon signed-rank.

Metric	p-value	Effect Type	Effect Size
Accuracy	0.813	Paired Cohen's d_z	0.258
Macro-F1	0.469	Paired Cohen's d_z	0.239
AUROC	0.813	Paired Cohen's d_z	0.097
EDV	0.016	Paired Cohen's d_z	-1.486
Instability	0.016	Paired Cohen's d_z	2.531
SSIM	9.54×10^{-7}	r	0.876

Table 4. Mean confusion matrices across 7 random seeds for the CNN and ViT models, with rows representing true classes, and columns representing predicted classes.

CNN Mean Confusion Matrix			
	Pred 0	Pred 1	Pred 2
True 0	206.71	1.14	2.14
True 1	0.43	178.43	31.14
True 2	7.86	55.71	125.43
ViT Mean Confusion Matrix			
	Pred 0	Pred 1	Pred 2
True 0	207.29	0.86	1.86
True 1	0.71	183.00	26.29
True 2	6.57	67.29	115.14

creasing percentages of top-k were removed, the CNN consistently exhibited larger performance drops at all masking levels, reducing from an accuracy of 0.568 at 10% removed (random = 0.790) to an accuracy of 0.320 at 90% removed (random = 0.299), compared to the ViT, which exhibited an accuracy of 0.863 at 10% removed (random = 0.867), and 0.380 at 90% removed (random = 0.488) (Table 3). These results suggest CNNs are more vulnerable to the targeting of their highest-saliency regions, indicating a stronger localized dependence on certain regions in determining a diagnosis, compared to the ViT, which exhibited a lower performance drop at the start. This suggests that ViTs rely more on global, scattered features in determining a diagnosis, and do not rely as significantly on localized features.

3.5. Explanation-Driven Vulnerability (EDV)

To assess the sensitivity of each model's explanations with targeted perturbations, explanation-driven vulnerability (EDV) was computed. A lower EDV value indicates more stable explanation behavior. Across 7 random seeds (4,263 samples), the CNN value was 0.0012 ± 0.0019, while the ViT value was 0.0003 ± 0.0006 (Table 1).

The ViT model had an $4\times$ reduction in its EDV value compared to the CNN, indicating a significantly greater robustness of its spatial explanations. These results suggest that, despite comparable predictive performance, transformer-based attention mechanisms might have a lower susceptibility to explanation collapse under perturbation.

4. Discussion

This study analyzed the reliability of explainable artificial intelligence for classifying keratoconus using corneal topography. Both reported comparable performances (CNN: 0.8384 ± 0.0580; ViT: 0.8299 ± 0.0431), enabling equal comparison. Due to their patch-level attention mechanisms, the diffuse appearance of the CAM maps generated from the ViT model aligns with prior research that Grad-CAM is less spatially precise for transformer-based architectures, meaning that CNNs rely more on localized regions, while ViTs use a broader spatial context when determining a conclusion.

The SSIM results (CNN: 0.4558 ± 0.1030, ViT: 0.6039 ± 0.1129) suggest that the ViT produces more consistent explanations across random seeds. Specifically, the higher cross-seed similarity suggests that vision transformers are able to learn more reproducible spatial corneal features, which can be used in clinical settings to improve trust in model explanations.

This analysis is further backed up by the instability results (CNN: 0.0013 ± 0.0004, ViT: 0.0001 ± 0.0001), with the CNN's generated explanations being significantly more unstable than those of the ViT (p = 0.016). The low EDV values for both the CNN and ViT models (CNN: 0.0012 ± 0.0019, ViT: 0.0003 ± 0.0006) suggest that, even with changes, the model predictions remained strong (Table 1). However, the ViT exhibited a large reduction in EDV compared to the CNN, suggesting a more stable behavior and reduced sensitivity to perturbations. These metrics both demonstrate that the CNN relies on localized features that could shift under small input changes, while the ViT relies on a spatial, global analysis for explanation generation.

However, a low EDV and higher CAM instability for the CNN show a disconnect between the stability of the model's predictions and the reliability of its explanations. While EDV indicates the sturdiness of the model's output probabilities under perturbation, CAM instability measures the sensitivity of spatial attribution maps. These results demonstrate a key problem with attempting to resolve the black box theory; even when CNNs have recurring predictions, their explanations vary significantly, making it more difficult for clinicians to comprehend how the machine learning

model reached the diagnosis compared to the ViT.

Removing the top-k regions highlighted by CAM resulted in different sensitivity patterns for the CNN and the ViT models. The CNN exhibited a significant drop in accuracy from 10% to 30% removal (0.568 to 0.349), with it eventually stabilizing after that. However, the ViT exhibited a mostly linear decrease from 10% to 70%, after which it plummeted at 90% removed (0.863 to 0.623 to 0.380). Interestingly, the ViT accuracy after significant top-k regions were removed was only slightly lower than its accuracy when random regions were removed, suggesting that faithfulness is architecture-dependent.

These results all suggest that CNN explanations are variable and unstable, but when targeted regions are removed, its accuracy drops quickly, meaning the CNN relies heavily on localized features. The ViT exhibited stable explanations and a gradual deletion effect, meaning that its features are globally distributed. This difference is likely because CNNs process images using spatially local convolutional kernels (which emphasize close pixel relationships and localized feature detection), while ViTs process images using patch tokens and use self-attention mechanisms, aggregating information globally.

5. Conclusion

This study investigated the reliability of explainable artificial intelligence (XAI) by comparing convolutional neural networks (CNNs) and vision transformers (ViTs) in corneal topography classification. We found that both the CNN and ViT had relatively high accuracies when trained on the independent test set - however, because performance alone is insufficient to assess clinical readiness, this study also evaluated the reliability of spatial explanations to determine the effectiveness of each model. Across all of the evaluated explainability robustness metrics, the ViT consistently exhibited demonstrated more stable and reproducible explanations than the CNN, serving as a benefit for clinicians focused on XAI explanation reliability.

Our deletion experiments found that CNN accuracy collapses when a small portion of regions classified by the model as "important" for a decision are removed, an effect only seen in ViTs after a significant percent of CAM regions are removed. As a result, the CNN most likely relies on more clustered spatial regions when developing a conclusion, and has a higher faithfulness than the ViT, although it has more unstable explanations.

In assessing keratoconus, clinicians using corneal topography rely on localized topographic features (like regions of corneal steepening or asymmetry). While the higher faithfulness of the CNN may make it more suited for keratoconus detection, its lower explanation stability compared to the ViT still means that the highlighted regions may vary between different explanations, reducing clinician confidence in the model's reasoning. Likewise, while the ViT has strong explanation stability, its reasoning derived from a global analysis may also not fully meet clinicians' needs when diagnosing keratoconus. These results suggest that further fine-tuning of CNNs and vision transformers are required before a full adoption in clinical settings.

Although we used ImageNet pretraining and data augmentation to address potential training limitations for transformer architectures from the modest dataset size (4,011 images), model generalization may still be constrained by the size of the dataset and how the clinical cases were distributed. Second, CAM-based methods were assessed as post-hoc explanations, rather than during the decision-making process itself, so there may have been instances when the model hallucinated "important" pixels when they were not significantly used in determining a diagnosis. This limitation may be most applicable for transformer-based architectures, because Grad-CAM can produce diffuse spatial maps due to the ViT having a global attention mechanism.

Thus, future studies should use different explainability frameworks (like SHAP, LIME, or Integrated Gradients) to address this issue. Also, using larger datasets and incorporating clinician-annotated regions of the images to check if the models correspond to clinically relevant regions of the cornea used for keratoconus detection would further evaluate the feasibility of CNNs and ViTs in keratoconus detection using corneal topography images.

Acknowledgments

No external funding was used for this study.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Cleveland Clinic. Corneal Ectasia: Causes & Symptoms, 2023. URL https://my.clevelandclinic.org/health/diseases/25178-corneal-ectasia.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://

arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Hammouch, E. Keratoconus detection. URL https://www.kaggle.com/datasets/elmehdi12/keratoconus-detection. Accessed: Dec. 19, 2025.

Jafarinasab, M. R., Shirzadeh, E., Feizi, S., Karimian, F., Akaberi, A., and Hasanpour, H. Sensitivity and specificity of posterior and anterior corneal elevation measured by orbscan in diagnosis of clinical and subclinical keratoconus. *Journal of Ophthalmic & Vision Research*, 10(1):10–15, 2015. ISSN 2008-2010. doi: 10.4103/2008-322X.156085.

Martínez-Abad, A. and Piñero, D. P. New perspectives on the detection and progression of keratoconus. *Journal of Cataract & Refractive Surgery*, 43(9):1213–1227, September 2017. ISSN 0886-3350. doi: 10.1016/j.jcrs.2017.07.021. URL https://www.sciencedirect.com/science/article/pii/S0886335017305527.

Muhsin, Z. J., Qahwaji, R., Ghafir, I., AlShawabkeh, M., Al Bdour, M., AlRyalat, S., and Al-Taee, M. Advances in machine learning for keratoconus diagnosis. *International Ophthalmology*, 45 (1):128, 2025. ISSN 0165-5701. doi: 10.1007/s10792-025-03496-4. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11955434/.

Nasim, M. A. A., Ferdous, A. S. M. A., Rashid, A., Soshi, F. T. J., Biswas, P., Biswas, A., and Gupta, K. D. Trustworthy XAI and Application, April 2025. URL http://arxiv.org/abs/2410.17139. arXiv:2410.17139.

Porter, D. Corneal Topography, May 2021. URL https://www.aao.org/eye-health/treatments/corneal-topography-4.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks?, March 2022. URL http://arxiv.org/abs/2108.08810. arXiv:2108.08810 [cs].

Santodomingo-Rubido, J., Carracedo, G., Suzaki, A., Villa-Collar, C., Vincent, S. J., and Wolffsohn, J. S. Keratoconus: An updated review. *Contact Lens and Anterior Eye*, 45(3):101559, 2022. URL https://www.contactlensjournal.com/article/S1367-0484(21)00205-8/fulltext.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, October 2017. doi: 10.1109/ICCV.2017.74. URL https://ieeexplore.ieee.org/document/8237336. ISSN: 2380-7504.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M., and Hamamoto, R. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *Journal of Medical Systems*, 48(1):84, September 2024. ISSN 1573-689X. doi: 10.1007/s10916-024-02105-8. URL https://doi.org/10.1007/s10916-024-02105-8.

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., and Wong, T. Y. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, February 2019. ISSN 0007-1161, 1468-2079. doi: 10.1136/bjophthalmol-2018-313173. URL https://bjo.bmj.com/content/103/2/167. Publisher: BMJ Publishing Group Ltd Section: Review.