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

Shankar Harikrishnan [*1]

1. Extended Abstract

This non-archival extended abstract summarizes work previously published in Pattern Recognition Letters: (Harikrishnan, 2026). The Pattern Recognition Letters publication is the archival version of this record.

While the use of explainable artificial intelligence (XAI) has become more common in medical imaging, the stability and reliability of its explanations remain unclear. We compare explanation behavior across computational neural networks (CNNs) and vision transformers (ViTs) in keratoconus diagnosis. We trained the models on 4,011 corneal topography images classified as normal, suspect, and keratoconus, and to assess reproducibility, we repeated training across 7 random seeds and generated Grad-CAM heatmaps for each trained model. Explanations were then evaluated with cross-seed heatmap consistency, structural similarity, instability of highlighted regions, sensitivity to CAM pixel removal, and explanation-driven vulnerability. Using these evaluation metrics allows us to distinguish between explanations that are visually stable across various training runs and explanations that are dependent on the model's decision behavior.

The results demonstrate that the explanation reliability significantly relies on model architecture. Specifically, we found that the ViT produced more consistent heatmaps across random seeds (cross-seed SSIM increasing from 0.456 in the CNN to 0.604 in the ViT), with there also being an order of magnitude reduction in explanation instability. However, the CNN exhibited a larger drop in confidence with CAM region removal, suggesting that the explanations were more sensitive to decision-relevant regions but also more variable across various runs. Overall, these results suggest a tradeoff between the explanation stability and faithfulness, as the ViT had stable but less faithful explanations, while the CNN had more faithful but more variable explanations.

This work represents a practical case study for evaluating the reliability various model architectures in keratoconus diagnosis with corneal topography images, suggesting that a further evaluation of visual explanations in XAI should be used before clinical deployment.

The full paper is available in the archival Pattern Recognition Letters publication: (Harikrishnan, 2026).

References

Harikrishnan, S. Instability and interpretability discrepancies between cnns and vision transformers in keratoconus detection. *Pattern Recognition Letters*, 204:107–112, 2026. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2026.03.020. URL https://www.sciencedirect.com/science/article/pii/S0167865526001108.

[1]Independent Researcher, Adlai E. Stevenson High School, Lincolnshire, Illinois, United States of America. Correspondence to: Shankar Harikrishnan <shankarhkn@gmail.com>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.