

# Developing an explainable artificial intelligence tool for training novices

Neil Barigye – Supervised by Dr Alina Bialkowski

[1] "HAM10000 expert-annotated explanations and pseudonymized reader study data." figshare. Accessed: Oct. 12, 2025. [Online]. Available: <https://figshare.com/s/c7feb070d066a4ccce19?file=56238653>  
[2] DBO-DKFZ/Derma-XAI. (Aug. 11, 2025). Jupyter Notebook. DBO-DKFZ. Accessed: Oct. 12, 2025. [Online]. Available: <https://github.com/DBO-DKFZ/Derma-XAI>  
[3] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "Help Me Help the AI: Understanding How Explainability Can Support Human-AI Interaction," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg Germany: ACM, Apr. 2023, pp. 1–17. doi: 10.1145/3544548.3581001.

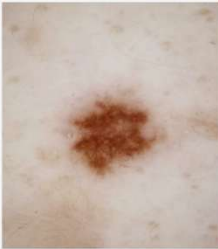
### Objective

This project aims to investigate whether explainable artificial intelligence (XAI) can be used to train novices to differentiate melanoma from nevi. If so, it examines which form of explanation is most effective.

### Method

Users participated in a two-phase study. In the first phase, they examined 40 skin lesions and classified each one as either a melanoma or a nevus (non-cancerous mole). The 40 trials were divided into four parts in a 6/14/6/14 sequence, where the 6-trial blocks were for training users and the 14-trial blocks were for testing understanding. The second phase of this study consisted of an 18-question (~5-minute) survey to gather their thoughts and feelings about the experience of their participation.

#### 1. ChatGPT Explanations



**Prediction:** Nevus  
**Confidence:** about 0.6 (moderate, not certain)

**Why:**

**Symmetry:** The pigmented area is roughly symmetric in shape and distribution around its center.

**Color:** Predominantly a single medium-brown tone with minimal variation; I don't see multiple distinct colors (e.g., black/blue/gray/white/red) that would raise suspicion.

**Border:** Although a bit fuzzy, the border is relatively even without prominent notches or peripheral streaks.

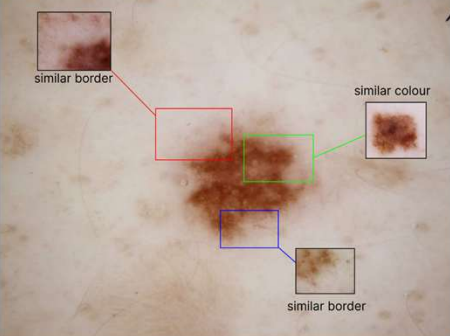
**Structures:** Largely structureless/homogeneous brown area without an atypical pigment network, irregular dots/globules at the periphery, negative network, blue-white veil, shiny white lines, or regression structures—features that would favor melanoma.

**Caveats:**

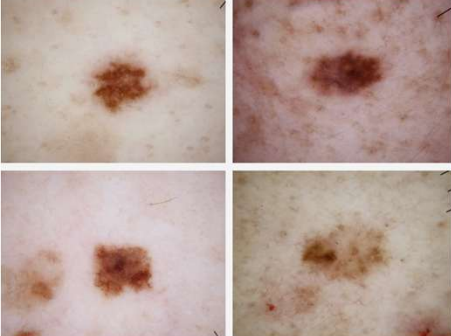
Size and evolution are not available; both are key (e.g., change over time, >6 mm).

Image focus/lighting can obscure subtle features.

#### 2. Prototype Explanations



#### 3. Example-Based Explanations



#### 4. Saliency Highlighting




Figure 1: Explanations produced by GPT-5 when asked to predict the status of the provided skin lesion, alongside a justification and confidence level [1].

Figure 2: Handmade mock explanations based on prototype explanations from [3], accompanied by a minimalist justification text (not shown) [1].

Figure 3: The four most visually similar lesions in the dataset (including duplicate images), accompanied by a minimalist justification text (not shown) [1].

Figure 4: Diagnostic areas of interest highlighted, accompanied by a minimalist justification text (not shown) [1], [2].

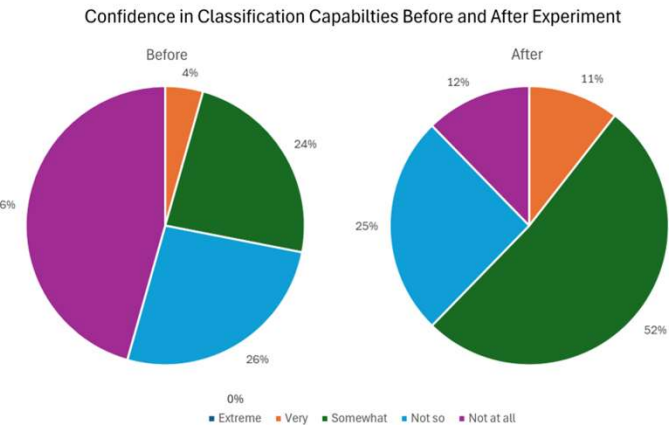


Figure 5: Users' confidence level in their ability to differentiate melanoma from nevi.

[n.barigye@uq.edu.au](mailto:n.barigye@uq.edu.au) Neil Barigye

School of Electrical Engineering and Computer Science

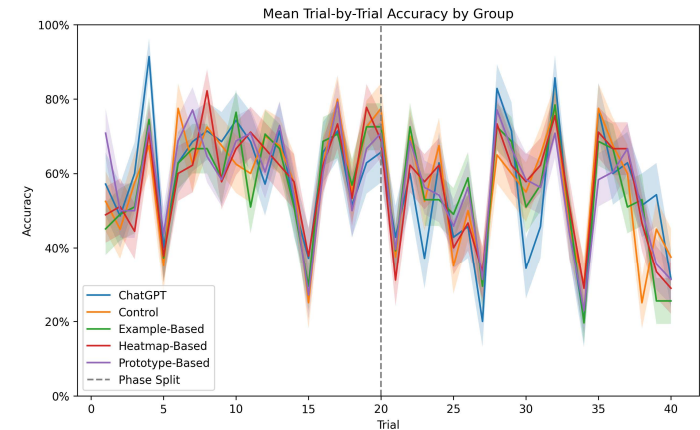


Figure 6: The mean trial-by-trial accuracy by group.

## Results & Conclusions

- Across all users and explanation types, performance followed an erratic trend.
- Although average performance across groups was roughly equivalent (56.5%, which represents fewer than three additional correct responses out of forty compared to random guessing), users consistently reported higher confidence in their ability to distinguish between melanoma and nevi.
- The experimental results suggest that, in this instance, explainable AI does not enhance novice training for melanoma-nevus classification beyond what can reasonably be achieved through trial and error or random guessing.