Interpretable AI for Medical Classification using ProtoPNet concept

Carlotta Hoelzle[1,2], Tuğcan Hoşer[1], and Hanad Abdullahi[1,3]

[1] Technical University of Munich, Munich, Germany
[2] KTH Royal Institute of Technology, Stockholm, Sweden
[3] Uppsala University, Uppsala, Sweden

Abstract. The integration of AI in medical imaging has significantly improved diagnostic capabilities, yet the interpretability of AI models remains a critical challenge for clinical application. This paper evaluates the performance of prototypical architectures in medical image classification and examines the utility of prototypical heatmaps in blood classification tasks. We assessed three ProtoPNet variations: ProtoPFormer, ProtoASNet, and PiPNet, on their accuracy and interpretability. Our results demonstrate that ProtoPFormer and ProtoASNet achieve performance on par with SOTA models, PiPNet exhibits notable interpretability without significantly compromising accuracy. The human study provides mixed feedback on heatmaps' usefulness, indicating areas for improving model explanations. This study aids in developing accurate, interpretable AI models, enhancing trust and usability in clinical settings.

Keywords: Interpretable AI · ProtoPNet · Human Study

1 Introduction

The integration of AI in healthcare, particularly medical imaging, has significantly enhanced diagnostic capabilities, offering higher precision, speed, and effectiveness in treatments. Nonetheless, a major challenge remains in the interpretability of AI models, crucial for clinical applicability. Current black-box models, despite their high performance, are opaque and their decision-making processes are not easily understood by medical professionals, leading to potential misinterpretations and oversight Habehh and Gohel [2021], Rudin [2019], Zech et al. [2018]. Post-hoc explanatory methods are used to analyze and interpret the models decision-making after it has made it's predictions, examples of these methods are saliency maps and GRAD-CAM Paul et al. [2023]. However, these approaches often produce misleading explanations that do not align with the models' internal logic Paul et al. [2023].
Intrinsic interpretability, where models are designed to be transparent from the outset, presents a solution. Chen et al. [2019] designed an intrinsic interpretable prototype network model, named ProtoPNet, which offers explainable classifications based on visually and semantically meaningful learned prototypes, providing clear reasoning for it's predictions Kumar et al. [2020]. This approach

could potentially aid medical professionals by demonstrating which image features lead to specific diagnostic conclusions, thereby enhancing trust and usability of AI recommendations in clinical decision-making. The core problem addressed in this research is balancing the accuracy and interpretability of AI models in healthcare. While black-box models excel in accuracy, their lack of transparency is problematic. Conversely, interpretable models like ProtoPNet, though transparent, often struggle to match the accuracy of black-box models due to the constraints imposed by their design for interpretability Gunning and Aha [2019]. This study aims to find a ProtoPNet architecture for medical image classification, maintaining high accuracy while providing interpretable and meaningful explanations. The research question investigates the performance of prototypical architectures in medical image classification and the utility of prototypical heatmaps in blood classification tasks. The paper is structured around the motivation for interpretable AI in healthcare, followed by a description of the methodology including the utilized architectures, datasets and settings, and a discussion of the quantitative and qualitative performances, as well as an analysis of the optimal number of prototypes per class regarding a specific dataset.

2 Related Work

Interpretability methods are broadly categorized into intrinsic (ante-hoc) and post-hoc. Intrinsic models are designed to be interpretable from the outset, embedding transparency within their architecture. In contrast, post-hoc methods explain predictions but often fall short in providing human-interpretable explanations, whereas ProtoPNet forces interpretability by linking input images to learned prototypes Kim et al. [2022]. Post-hoc methods like salience maps or Grad-CAMs highlight important regions of input images for the prediction but can produce explanations that do not fully align with the model's internal logic, leading to potential misinterpretations in critical medical decisions Paul et al. [2023], Doshi-Velez and Kim [2017], Qi et al. [2023]. Furthermore, these explanations tend to be limited in scope, potentially causing over-reliance or under-reliance on the AI system by the users Lage et al. [2019]. Several post-hoc interpretable models have been used in medical imaging, e.g. Locally Interpretable Model-Agnostic Explanations (LIME) and DeepLIFT. LIME uses surrogate models to provide localized insights into individual predictions, offering flexibility across different imaging tasks, but its approximations can be unstable and inconsistent Wu et al. [2020], Doshi-Velez and Kim [2017]. DeepLIFT dissects neural network outputs by backpropagating contributions of each input feature, providing detailed attribution maps; however, it can produce misleading explanations, which is problematic in clinical contexts Ennab and Mcheick [2022], Kim et al. [2022]. Ante-hoc methods, such as linear regression and decision trees, offer decision-making processes that are transparent and more interpretable compared to black-box models Wu et al. [2020]. However, these models often struggle to maintain high predictive accuracy for some tasks compared to more complex models, and their simplicity may not capture intricate patterns

in the data, limiting their effectiveness in visual medical applications Ennab and Mcheick [2022].
ProtoPNet by Chen et al. [2019] is a deep-learning ant-hoc method that uses prototypes to make decisions, providing an interpretable framework that aligns with human reasoning. The architecture learns prototypes during training which are then used to make predictions. Visualization of these prototypes provide clear visual and semantic cues that help users understand why certain decisions are made Lage et al. [2019]. Unlike post-hoc methods, the semantic cues are directly tied to its decision-making process, posing the possibility of usability in clinical settings Paul et al. [2023]. However, the literature states that while ProtoPNet provides a high level of interpretability, maintaining this without compromising on predictive accuracy remains a significant challenge Gunning and Aha [2019]. This paper will investigate whether prototypical networks, networks that utilizes case-based reasoning like the ProtoPNet, are able to achieve the accuracy of state-of-the-art methods and if their interpretability outputs are relevant for clinicians. For this, we will use three different approaches of the ProtoPNet concept, two based on convolutional networks and one employing a transformer approach.

3 Methodology

This section outlines the architectural comparisons, datasets, and experimental setups utilized in our study.

3.1 ProtoPNet Architectures

The authors compared multiple published ProtoPNet variations to identify three top-performing models for downstream qualitative and quantitative performance analysis. The networks selected from the literature included ProtoPNet, ProtoPDebug, ProtoSeNet, ProtoPFormer, ProtoASNet, Deformable ProtoPNet, ProtoPool, ProtoTree, and PiPNet. To determine the most suitable architectures, the authors tested each model on the PriMIA dataset, considering only those with a mean test Matthews Correlation Coefficient of 74.4% or above. This testing criterion left three networks: ProtoPFormer, ProtoASNet, and ProtoPiPNet. The ProtoPFormer, a transformer-based model, is compared to the SOTA transformer model MedViT-T Manzari et al. [2023]. ProtoASNet and PiPNet, both using a ResNet50 backbone, are compared to a standard ResNet50. ResNet50 and MedViT-T were chosen for their top performance and comparable architecture. These top-performing variations will be quantiatively and qualitatively analyzed and compared to assess their performance and utility in medical classification tasks. **ProtoPFormer**, introduced by Xue et al. [2022], is a variation of ProtoPNet that uses a Vision Transformer architecture in place of the conventional convolutional network. This model employs an ImageNet pre-trained DeiT_tiny backbone, consisting of about 5.6 million parameters, 5 million fewer than the SOTA model ProtoPFormer integrates both global and

local information within images, by leveraging self-attention mechanisms and using two distinct prototype branches. The prediction of each branch is weighted to contribute to the final classification decisions. Local prototypes are directed towards capturing heterogeneous visual parts and foreground elements, whereas global prototypes focus on the overall image context. In ProtoPFormer, the optimization process centers around refining a feature map to represent prototypes, rather than directly learning image patches. This map is iteratively updated via backpropagation, allowing the model to refine prototypes during training. To translate learned prototype feature maps back into image space, a new method was devised. This method iterates through training images to identify patches that closely match the learned feature map. Alternatively, an encoder-decoder network could be trained to synthesize new prototypes within the image space from the learned feature map. **ProtoASNet** by Vaseli et al. [2023] introduces ProtoASNet, an advanced ProtoPNet variant using Convolutional Neural Networks (CNNs). It features a ResNet50 backbone with pre-trained ImageNet weights, totaling around 29M parameters. ProtoASNet captures spatial patterns like textures, edges, and shapes through its CNNs, managing static data effectively. It incorporates aleatoric uncertainty estimation by adding trainable uncertainty prototypes, which quantify prediction uncertainty by comparing the feature representation of the input with these prototypes. ProtoASNet helps avoid unreliable predictions in regions of high aleatoric uncertainty, aiding better decision-making. The model also identifies Regions of Interest within feature maps, comparing them to learned prototypes for similarity scores. An "abstention loss" method is used to learn the uncertainty prototypes, encouraging the model to abstain from uncertain predictions by penalizing overconfident incorrect predictions. This quantification of uncertainty enhances clinical decision-making by providing confidence measures, prioritizing high-certainty diagnoses, and detecting poor-quality data.

Nauta et al. [2023] introduce **PIPNet**, a CNN-based ProtoPNet version with a ResNet50 backbone and a total of 24M parameters. The architecture uses a CNN to learn prototypical representations pooled into prototype presence scores. Contrastive learning and a tanh-loss function ensure consistent prototype assignment and regularize the model. A sparse linear layer links prototypes directly to classification, ensuring faithful explanations and reducing the chance of misleading interpretations. The main difference between PipNet and ProtoPNet is that PIPNet focuses on precise part-based localization, such as keypoint detection, by processing image parts to improve accuracy in identifying specific image features. PIPNet also automatically identifies image parts that match human visual perception, using a sparse linear layer to connect prototypes to classes. This layer acts as a scoring sheet. The model combines supervised and self-supervised learning, starting with pretraining that uses alignment and uniformity losses, followed by fine-tuning for classification. Prototypes are visualized as image patches, offering both global and local explanations of decisions. PIPNet uses only image-level labels, requiring no part annotations. It is designed to be intuitive, compact, and capable of handling out-of-distribution data, ensuring usability and robustness.

PIPNet achieves this by utilizing a sparse linear decision layer that allows it to avoid making a decision when relevant prototypes are not present, effectively identifying unfamiliar samples.

3.2 Datasets

We used a three-class X-Ray dataset for the initial selection of ProtoPNet networks. For further evaluation, we utilized the refined MedMNIST with 224x224 pixel images, assuming that the higher resolution and image region of interest may allow the model to derive more meaningful prototypes. This dataset enabled comparison of each ProtoPNet against SOTA models, testing performance on multi- and binary class, and RGB and BW datasets, using the predefined training and evaluation splits. To assess robustness and generalizability, a second real-world Ultrasound dataset was used, offering insights into network performance in more varied clinical scenarios.

3.3 Experimental Setup

The full implementation of our experiments can be found on this GitHub. SOTA results were sourced from this review. The ProtoPFormer, ProtoASNet, and PiPNet were trained on a NVIDIA TITAN Xp, a NVIDIA Quadro RTX 6000, and a NVIDIA A100 Tensor Core GPU, respectively. All networks were trained with a batch size of 64 for 100 epochs, using ImageNet normalization without data augmentation for all datasets. The mean ACC and AUC test set scores over 5 runs were computed to provide robust performance estimates. Visual performance was evaluated by constructing heatmaps of the three most influential prototypes per class for every image, weighted according to their contribution to the class prediction. Heatmaps of three classes for two images were used in a human study to receive qualitative feedback from medical experts.

4 Discussion

This section provides a comprehensive analysis of the quantitative and qualitative evaluations of the ProtoPNet variants and their comparison with state-of-the-art models.

4.1 Quantitative Evaluation

The quantitative evaluation involved comparing the accuracy (ACC) and area under the curve (AUC) of the three prototypical networks PiPNet, ProtoPFormer, and ProtoASNet, each using 10 prototypes per class, against each other and state-of-the-art models. We employed two different SOTA models: a classic CNN with a ResNet-50 backbone, and a Vision Transformer. The results of the quantitative evaluation are presented in Figure 1. PiPNet underperformed 5-10% on all MedMNIST datasets for AUC and ACC, with the exception of

the BloodMNIST and PneumoniaMNIST datasets. A plausible reason for this could be PIP-Net's architectural focus on reducing the semantic gap between the pixel space and the latent space, which might result in a greater loss of accuracy. In contrast, both ProtoPFormer and ProtoASNet demonstrated comparable performance to the SOTA models, with ProtoPFormer outperforming ProtoASNet on all datasets on AUC. Specifically, both networks outperformed each related SOTA model on a majority of MedMNIST datasets, with the exceptions of BreastMNIST and PathMNIST for ProtoPFormer, and ProtoASNet on PneumoniaMNIST. However, fine-tuning the number of prototypes per class to 3 and 5 for BreastMNIST and PneumoniaMNIST, respectively, resulted in ProtoPFormer outperforming the SOTA Vision Transformer again. } No clear trend emerged between the three prototypical networks and their comparison with the SOTA models. The superior performance of ProtoPFormer may be attributed to its unique approach of optimizing a feature prototype vector instead of inherently learning image prototypes. This approach offers greater flexibility and potentially higher accuracy, as it is not constrained by the need to match learned image patches directly to the images they predict. ProtoASNet's superiority may be due to incorporating aleatoric uncertainty estimation and 4 million additional parameters compared to the SOTA model, though this hypothesis requires further validation.

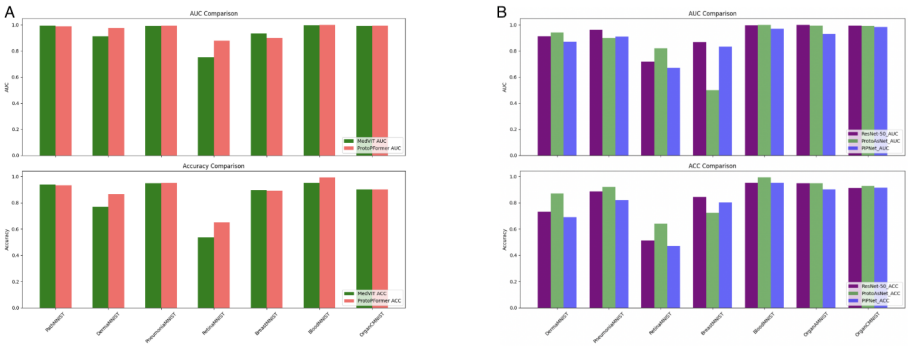

Fig. 1. The image represents the full quantitative comparison made between ProtoPformer, PIPNet, PIPNet and SOTA. **A)** represents the comparison made with transformer backbone. **B)** represents the comparison made with a CNN backbone.

4.2 Prototype Analysis

In our quantitative evaluation, we also investigated the effect of the number of prototypes per class on the performance of ProtoPFormer, ProtoASNet, and PiPNet. This analysis was conducted across datasets with different characteristics, including multi-class versus binary classification and RGB versus black-and-white images. Across ProtoPFormer and PiPNet, a common pattern emerged:

for binary classification tasks, the use of three prototypes generally provided optimal performance. In contrast, multi-class classification tasks benefited from a higher number of prototypes, with five or ten prototypes consistently yielding better results. Additionally, for BW images, a smaller number of prototypes, such as three, was often sufficient, whereas for RGB images, a larger number, such as five or ten, improved performance. For ProtoASNet, no clear trends were observed. The only exception was that for multi-class tasks and RGB images using ten prototypes per class never yielded the best performance, which contradicts the findings from the other networks. This analysis shows that the optimal number of prototypes per class can greatly depend on the specific characteristics of the dataset. Tailoring the number of prototypes accordingly can enhance model performance.

4.3 Qualitative Evaluation

ProtoASNet was excluded from the qualitative evaluation due to consistently producing identical activations for every prototype, as shown in Figure 2. This indicates the network's failure to learn meaningful prototypes. Following discussions with medical students, we decided to overlay the three most influential prototypes per class into a single heatmap rather than using individual prototype boxes, see Figure 3 for comparison. This decision was based on the students' familiarity with heatmaps and the need to consolidate information to prevent overload. The top three prototypes per class were weighted by their influence on the final prediction and then combined into one heatmap per class for each image. To evaluate the usability of prototypes in the medical domain, we conducted a human study involving one medical expert and five medical students using the BloodMNIST dataset. This dataset was chosen for its relatively simple, less dependent on high resolution, multi-class classification task. The study included an introduction to prototypical networks and heatmaps, followed by questions on participants' knowledge of hematology and AI in medical classification. The second part aimed to assess the utility of the prototypes for clinicians. The questionnaire is detailed in Appendix A. Participants first completed an objective evaluation task by selecting the predicted class from three possible classes to determine if the visualized prototypes are meaningful in distinguishing between predicted and non-predicted classes. The subsequent subjective evaluation involved assessing the meaningfulness of prototypes for correctly classified classes. This was done by showing participants the original image and the heatmap with predicted prototypes, asking if the focuses were correct. The final subjective question asked participants to choose the preferred heatmap of all three models. These evaluations provided insights into the qualitative performance and usability of the models in a medical context. In the objective task, users could not consistently identify the correct class predicted by the models. When an incorrect class was selected for one network, it was often the same incorrect class for other networks, suggesting underlying biases. The subjective evaluation of prototype meaningfulness yielded diverse results. Most participants found ProtoPFormer prototypes partially correct for one image and unclear for another. For PiPNet,

responses ranged from unclear to correct for one image and incorrect to partially correct for another, indicating that PiPNet prototypes were more helpful in distinguishing correct from incorrect focuses. The final subjective question reinforced these findings, with the majority of participants preferring PiPNet for assigning the model's focus in image classification.

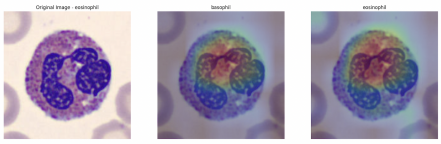

Fig. 2. ProtoASNet Heatmap for 2 Classifications

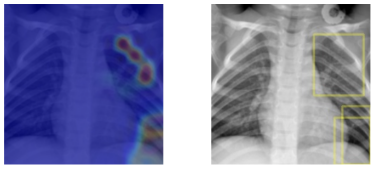

Fig. 3. Top 3 Normal Chest X-Rays: Heatmap & Bounding Boxes Overlay

5 Conclusion

This study explored the balance between accuracy and interpretability in prototypical architectures for medical image classification, with a particular focus on blood classification tasks. Our findings indicate that ProtoPFormer and ProtoASNet can achieve comparable accuracy to SOTA models, while PiPNet offers enhanced interpretability with a minimal trade-off in performance. The human study provided valuable insights, revealing that while prototypical heatmaps make AI models more interpretable, their utility varies among users. Medical professionals appreciated the intuitive visual explanations, but inconsistencies in prototype quality and clarity underscored the need for further refinement. Overall, this research demonstrates the potential of prototypical networks to bridge the gap between performance and interpretability in medical AI. Future work should enhance prototype clarity and consistency and expand the evaluation to include a broader range of medical tasks and experts. By refining these models, we can move closer to AI systems that are both highly accurate and transparently interpretable, ultimately improving their integration into clinical practice.

Bibliography

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Mohammad Ennab and Hamid Mcheick. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. *Diagnostics*, 12(7):1557, 2022.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

Hafsa Habehh and Suril Gohel. Machine learning in healthcare. *Current genomics*, 22(4):291, 2021.

Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pages 5491–5500. PMLR, 2020.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157: 106791, 2023.

Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pipnet: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2744–2753, 2023.

Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Carlyn, Samuel Stevens, Kaiya Provost, Anuj Karpatne, Bryan Carstens, Daniel Rubenstein, et al. A simple interpretable transformer for fine-grained image classification and analysis. *arXiv preprint arXiv:2311.04157*, 2023.

R Qi, Y Zheng, Y Yang, CC Cao, and JH Hsiao. Explanation strategies for image classification in humans vs. current explainable ai. arxiv, 2023.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. nat mach intell 1: 206–215, 2019.

Hooman Vaseli, Ang Nan Gu, S Neda Ahmadi Amiri, Michael Y Tsang, Andrea Fung, Nima Kondori, Armin Saadat, Purang Abolmaesumi, and Teresa SM

Tsang. Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 368–378. Springer, 2023.

Junjie Wu, Guannan Liu, Jingyuan Wang, Yuan Zuo, Hui Bu, and Hao Lin. Data intelligence: Trends and challenges. *Syst. Eng.-Theory Pract*, 40:2116–2149, 2020.

Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A Supplementary Material

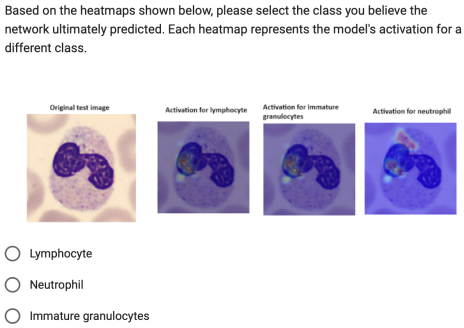

Fig. 4. Objective Evaluation: Distinction Study

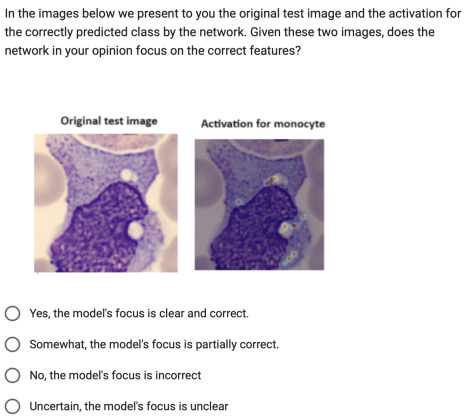

Fig. 5. Subjective Evaluation: Meaningful Prototypes

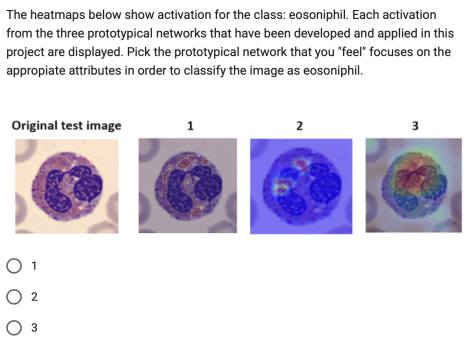

Fig. 6. Subjective Evaluation: Preferred Model Prediction