How effective and interpretable are ProtoPNet architectures in medical image classification, and how can the use of prototypes be optimized to aid clinical decision-making?

Firstname1 Lastname1 [*1] **Firstname2 Lastname2** [*1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3] **Firstname5 Lastname5** [1] **Firstname6 Lastname6** [3 1 2] **Firstname7 Lastname7** [2] **Firstname8 Lastname8** [3] **Firstname8 Lastname8** [1 2]

Abstract

Briefly summarize the motivation, research question, methodology, key results, and implications.

1. Introduction

The integration of Artificial Intelligence (AI) in healthcare, particularly in medical imaging, has significantly advanced the capabilities of medical diagnostics. Machine learning (ML) technologies have enabled higher precision, speed, and effectiveness in treatments and diagnostic processes, augmenting the abilities of physicians and specialists (Habehh & Gohel, 2021). For instance, areas such as radiology and in-hospital patient monitoring have seen notable advancements due to AI's ability to process large volumes of data rapidly and accurately. However, the major challenge in deploying these technologies is not just in their capacity to perform but in their interpretability and the clinical applicability of their results (Rudin, 2019).

Despite these advancements, the current state of AI in healthcare often involves the use of black-box models whose decision-making processes are opaque and not easily understandable by medical professionals. This lack of transparency can lead to misinterpretations and oversight, such as failing to recognize biases in data or errors in feature recognition—an issue evident when a neural network mistakenly identified the word 'portable' as a clinical indicator in X-ray images (Zech et al., 2018). Thus, there's a clear gap in the clinical usability of AI outputs due to the insufficient interpretability of these systems. The complications arising from these models have led to the use of post-hoc explanatory methods such as saliency maps and GRAD-CAM. However, these methods often yield explanations that are not only potentially misleading but also detached from the actual computations and logic of the underlying models not necessarily aligned with how models make predictions (Paul et al., 2023). For instance, saliency maps, while highlighting influential image regions, fail to explain how these regions affect the model's decisions, often producing similar visual explanations for distinctly different outputs (Rudin, 2019).

In contrast, intrinsically interpretable models are designed to be transparent from the outset, embedding explainability directly within their architecture. By using methods such as prototype networks, these models base their decisions on understandable and visually identifiable components that are aligned with domain-specific knowledge and constraints, such as causality, structural rules, and physical laws (Kumar et al., 2020). The explanations provided by these models are direct reflections of their computational processes, ensuring that the interpretative outputs are both accurate and actionable within the healthcare context. ProtoPNets are designed to make classifications based on learned prototypes that are visually and semantically meaningful. This type of model can aid medical professionals by providing transparent reasoning for each diagnosis. For example, in medical imaging, a ProtoPNet can demonstrate which features of an image led to a particular diagnostic conclusion, such as highlighting specific patterns in an X-ray that correspond to learned prototypes indicative of a disease. This not only aids in the diagnostic process but also empowers healthcare providers to understand, trust, and effectively utilize AI recommendations in their decision-making processes.

One of the perennial challenges in the deployment of AI in healthcare is balancing the accuracy of machine learning models with their interpretability. While black-box models often achieve high accuracy, their lack of transparency can be a significant drawback in clinical settings where understanding the 'why' behind a diagnosis is as critical as the diagnosis itself. Conversely, while interpretable models like

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

ProtoPNets provide transparency, ensuring they match the accuracy of the best-performing black-box models remains a challenge (Gunning & Aha, 2019). Furthermore, how the interpretability of the model is presented to the healthcare expert to help and not further overwhelm remains an open question. This research focuses on the optimization of interpretable ProtoPNets for medical image classsification and the goal of maintaining high levels of accuracy while also providing a framework how to provide clear, actionable, and clinically relevant explanations.

2. Related Work / Background

- Explain prototype-based learning and its significance. argue that the uncertainty in question potentially undermines the epistemic authority of clinicians. (Grote & Berens, 2020)

- Describe the original ProtoPNet architecture. Inspired by ProtoPNet, ProtoTree (Nauta et al., 2021) arranges the comparison to prototypes in a tree structure to mimic human reasoning; ProtoPFormer (Xue et al., 2022) presents a Transformer-based realization of ProtoPNet, which was originally based on ConvNets. Along with these interpretable decision processes, however, come specifically tailored architecture designs and increased complexity of the training process, often making them hard to reproduce, adapt, or extend. For instance, ProtoPNet requires a multi-stage training strategy, each stage taking care of a portion of the learnable parameters including the prototypes (Paul et al., 2023) other protoPNets with transformer: none ProtoP-OD: Explainable Object Detection with Prototypical Part only network found which has a transformer decoder but still uses a ResNet Backbone as encoder

- Discuss the equivalence condition for interpretability (ProtoPNet vs. human similarity).: In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human. (Doshi-Velez & Kim, 2017) In the context of local explanations for predictions madeby ML models, the phenomenon to be explained is why amodelMpredicted outputyfor inputx. The facts used toexplain this phenomenon may include information about theinput features, the model parameters, the data used to trainthe model, or the manner in which the model was trained. Yet, the expla-nations produced by most current interpretability methodsrefer only to why the inputx-points to a single hypothesis(i.e., the predictiony) rather than ruling out all alternatives. 1. Explanations should be contrastive, i.e., explicate whythe model predictedyinstead of alternativey.2.Explanations should be exhaustive, i.e., provide a jus-tification for why every alternativeywas not predicted 3. Explanations should be modular and compositional,breaking up predictions into simple components 5.Explanations should be parsimonious, i.e., only themost relevant facts should be provided as components. **Explaining with the weight of evidence**Weoperationalize the question of why modelMpredicted out-putyfor inputxin terms of how muchevidenceeach inputfeaturexi(or feature group) provides in favor ofyrelative toalternatives. -.¿ grounded in common language,it naturally evokes a contrastive statement (evidencefororagainstsomething) (Melis et al., 2021) what kinds of explanation are truly human-interpretable remains poorly understood three main fields of what makes explanations interpretable related to the tasks that users may perform with machine learning systems: simulation of the response, verification of a suggested response, and determining whether the correctness of a suggested response changes under a change to the inputs. (Lage et al., 2019) For example, humans prefer explanations that are both simple and highly probable [Lombrozo, 2007]. Miller [1956] famously argued that humans can hold about seven items simultaneously in working memory, suggesting that human-interpretable explanations should obey some kind of capacity limit **Examples as Representations:** Humans perform better when they can relate abstract concepts to concrete examples. This method allows for a more intuitive understanding by linking new information to familiar instances, thereby making the decision-making process more relatable and understandable. (Kim et al., 2014) **Decision Sets:** Decision sets provide a concise and clear set of if-then rules that describe how decisions are made. This format helps users quickly understand the logic behind decisions, as they can easily trace the reasoning process without dealing with the complexity of intertwined rules, enhancing both speed and accuracy in decision-making. (Lakkaraju et al., 2016) **Decision Trees:** Decision trees offer a visual breakdown of decision processes, presenting a clear, branching structure of choices and outcomes. This method appeals to humans as it visually simulates the decision-making process, allowing for easier comprehension and analysis of how different choices lead to different consequences. (Subramanian et al., 1992) Horsky et al. [2012] describe how presenting the right clinical data alongside a decision support recommendation can help with adoption and trust Bussone et al. [2015] found that overly detailed explanations from clinical decision support systems enhance trust but also create over-reliance; short or absent explanations prevent over-reliance but decrease trust.

"Having the ability to describe or express something in terms that a human can understand" is one of the most common definitions of interpretability. "The level to which a person recognizes the reason for a selection" is another standard definition

- Review previous work on interpretability in AI.

The model is considered inherently interpretable if a person can comprehend its underlying workings, either the complete model at once or at least the elements of the model

relevant to a specific prediction In contrast, we consider the model's prediction explainable if a process can offer (partial) knowledge about the model's workings.(Ennab & Mcheick, 2022) However, despite the interest in interpretability, there is very little consensus on what interpretable machine learning is and how it should be measured.(Doshi-Velez & Kim, 2017) Interpretable machine learning methods aim to optimize models for both succinct explanation and predictive performance. Common types of explanation include regressions with simple, human-simulatable functions [Caruana et al., 2015, Kim et al., 2015a, Ruping, 2006, Bucilu et al., 2006, Ustun and Rudin, 2016, Doshi-Velez et al., 2015, Kim et al., ¨ 2015b, Krakovna and Doshi-Velez, 2016, Hughes et al., 2016, Jung et al., 2017], various kinds of logic-based methods [Wang and Rudin, 2015, Lakkaraju et al., 2016, Singh et al., 2016, Liu and Tsang, 2016, Safavian and Landgrebe, 1991, Wang et al., 2017], techniques for extracting local explanations from black-box models [Ribeiro et al., 2016, Lei et al., 2016, Adler et al., 2016, Selvaraju et al., 2016, Smilkov et al., 2017, Shrikumar et al., 2016, Kindermans et al., 2017, Ross et al., 2017], and visualization [Wattenberg et al., 2016] As surveyed in (Zhang Zhu, 2018; Burkart Huber, 2021; Carvalho et al., 2019; Das Rad, 2020; Buhrmester et al., 2021; Linardatos et al., 2020), various ways exist to explain or interpret a model's prediction (see Appendix A for more details). Among them, the most popular is localizing where the model looks for predicting a particular class.

methods need to have multiple reports per image or with false positives on the same image as a true positive and/or false negative (miss) Color displays are becoming an important modality(Krupinski, 2010) A common XAI method for image classification has been using saliency maps that highlight regions of the input image according to their importance to the AI model's output (Li et al., 2021) Two major saliency-map based approaches are perturbation-based and backpropagation-based methods. Perturbation-based methods, such as RISE (Petsiuk et al., 2018), perturb the input image and place more weights on the pixels that affect the output class probability relative to other classes when occluded. In contrast, backpropagation-based methods, such as GradCAM (Selvaraju et al., 2020), calculate the gradient of the score for the target class in a particular layer as the class relevance of each pixel. These saliency maps have often been compared with human attention maps under the assumption that humans attend to features important to their judgements during image classification. saliency-map based XAI highlights image regions that contribute to the classifier output, and thus should be compared with human attention when performing image classification tasks. Understanding whether saliency maps generated using the current XAI methods are better matched with human attention during image classification or explanation will provide important insights on what information these XAI salience maps reflect and how human users should interpret them.suggesting that focusing on identifying critical features of the foreground object is beneficial for classification (Qi et al., 2023)

Ante-hoc is another name for the transparent or intrinsic approach. A model that can be understood independently is said to be transparent. Intrinsic explainability can be attained by creating self-explanatory models that explicitly represent meaningfulness in their design. When explanation techniques are used after model training, it is called post-hoc explainability Post-hoc explana- tions focus on explaining predictions made by already-trained models, whereas interpretable-by-design (IBD) models are intentionally designed to possess a ... (Kim et al., 2022)

Understanding a model's overall decision logic is called global explainability, whereas local explainability is concerned with the justifications of specific predictions.

Participants prefer to use a model with explanations over a baseline model without explanations. To switch their preference, they require the baseline model to have +6.2% to +10.9% higher accuracy. (Kim et al., 2022)

ever, recent works suggest that some methods are not as interpretable as originally imagined and may engender overtrust in automated systems (Kim et al., 2022)

- Discuss other interpretable models used in medical imaging. interpretable model (such as linear regression or a decision tree) Locally Interpretable Model-Agnostic Explanations (LIME) surrogate model that uses a trained local model to interpret a single sample (Wu et al., 2020) DeepLIFT is a method for dissecting the output prediction of the neural network on a given input by backpropagating the contributions of all neurons in the network to each characteristic of the input. DeepLIFT assigns value to neurons depending on their activity. When the local gradient is zero, the findings might be deceptive. DeepLIFT produces surprisingly distinct attribution maps from input CT images with minor perturbations that are visually identical. Almost all of these strategies aim towards local explainability or justifying decisions in a particular case (Ennab & Mcheick, 2022)

Interestingly, XAI saliency-map explanations had the highest similarity to the explorative attention strategy in humans, and explanations highlighting discriminative features from invoking observable causality through perturbation had higher similarity to human strategies than those highlighting internal features associated with higher class score. Several studies investigated the fusion of image captioning with visual explanations, as shown in Fig. 4, called Image Captioning with Visual Explanations (Xu et al., 2015) Case-based explanations, textual explanations, and auxiliary explanations.(Borys et al., 2023) automatic metrics poorly

correlate with human perfor- mance in post-hoc attribution heatmap evaluation

- Compare ProtoPNet with these models in terms of performance and interpretability. which regions are actually responsible for the prediction - Comparative analysis highlighting where ProtoPNet stands out or falls short.

- how do humans reason / make decsion / when do they trust the model / Human interpretability A human experiment needs to be well-designed to minimize confounding factors, consumed time, and other resources. the best way to show that the model works is to evaluate it with respect to the task: doctors performing diagnoses

When making a decision, there is so much potentially relevant information available, it is impossible to know or process it all (so called 'bounded rationality') incorporate the findings from high quality research into routine clinical practice (Grote & Berens, 2020)

truncating the amount of information used in order to be able to make a 'good enough' decision relevant information available to a decision maker that it is impossible for the human brain to know or process it all(Simon, 1979)

Need to know how "sure" the AI is with it's decision (Ennab & Mcheick, 2022)

that explanations engender human trust, even for incorrect predictions, yet are not distinct enough for users to distinguish between correct and incorrect predictions (Kim et al., 2022)

An MIT study from 2014 found that people can identify an image in as fast as 13 milliseconds Thus, when people view stimuli for 50 ms or less with backward pattern masking, as in some conditions in the present study, the observer may have too little time for reentrant loops to be established between higher and lower levels of the visual hierarchy before earlier stages of processing are interrupted by the subsequent mask (Potter et al., 2014)

typically involve directing attention to relevant details following a sequence of visual reasoning processes, in contrast to XAI methods that simply highlight features used by AI classifiers without temporal information people tend to provide contrastive explanations that focus on why the current event occurs instead of other non-occurring events making decisions based on complex perceptual processes that are often automatic and unconscious in humans such as image classification (Gupta & Seeja, 2024)

confirmation bias: when provided explanations, participants tend to believe that the model predictions are correct (Kim et al., 2022)

- Human and Machine Aspects of Trust: context of use (who, why, when, where) can help answer target questions about explanation design choices such as what information the explanation needs to contain (i.e., the content) and how that information needs to be provided who: experts why: when: where: hospital setting / doctors office what: local for one patient global to establish trust in the system how:

3. Methodology

3.1. ProtoPNet Architectures

- Describe the different ProtoPNet variants evaluated (ProtoPFormer, ProtoPool, ProtoASNet, PIPNet). - Highlight key features and differences between these architectures. - Justification for selecting these architectures for evaluation. - Binary forced choice: humans are presented with pairs of explanations, and must choose the one that they find of higher quality (basic face-validity test made quantitative). - Forward simulation/prediction: humans are presented with an explanation and an input, and must correctly simulate the model's output (regardless of the true output). - Counterfactual simulation: humans are presented with an explanation, an input, and an output, and are asked what must be changed to change the method's prediction to a desired output (and related variants). (Doshi-Velez & Kim, 2017) g. In practice, the joint presentation of the radiographic findings and inferential analysis at the start provides a more comprehensive set of information sources to the decision maker. However, as AI inferences may be erroneous, there are concerns about the possibility of unwanted anchoring that could lower the team's performance human-AI collaborations for optimal team decision making human-centered design need to be considered A handful of studies have compared the influences of various types of AI assistance on decisions At the same time, some of these analyses have also witnessed the pitfalls of AI adoption, including increases in the time spent on the task without corresponding gains in accuracy, reductions in diagnostic performance due to reliance on erroneous AI advice, and participants ignoring AI advice altogether (Fogliato et al., 2022) Decision making by health care professionals is often complicated by the need to integrate ill-structured, uncertain, and potentially conflicting information from various sources in order to build information systems that can support complex decision making it will be necessary to more fully understand human decision-making processes (Kushniruk, 2001) Physicians are working longer hours than ever before, and concerns have been raised regarding fatigue and whether it adversely affects diagnostic accuracy. (Krupinski, 2010)

3.2. Datasets

- Detail the medical datasets used (Ultrasound, MedMNIST). - Discuss their characteristics and why they were chosen. - Preprocessing steps and any augmentation techniques used.

3.3. Experimental Setup

- Explain the evaluation metrics (ACC, AUC, MCC). - Describe the experimental procedure for testing ProtoPNet architectures (training and evaluation) - Outline the prototype variation experiment (3, 5, 8 prototypes per class). - Discuss the method for expert evaluation of prototypes. However, these automatic evaluation metrics are dis- connected from downstream use cases of explanations; they don't capture how useful end-users find heatmaps in their decision making. (Kim et al., 2022) Research in interpretability inherently faces the challenge of effective and reliable validation (Doshi-Velez & Kim, 2017) Understanding a model's overall decision logic is called global explainability, whereas local explainability is concerned with the justifications of specific predictions. Global: learned prototypes Local: inference with distance to learned prototypes and weight vectors to prototypes from other classes (Gupta & Seeja, 2024)

HIVE to the best of the authors knowledge the only framework to human evaluation framework that assesses the utility of explanations to human users in AI-assisted decision mak- ing scenarios, and enables falsifiable hypothesis testing (evaluate how useful explanations are in identifying errors made by a model.) , cross-method comparison, and human-centered evaluation of visual interpretability methods. A falsifiable hypothesis is one that can be proven false through evidence (Kim et al., 2022)

When given multiple model predictions and explanations, participants struggle to distinguish between correct and incorrect predictions based on the explanations. This result suggests that interpretability methods need to be improved to be reliably useful for AI-assisted decision making.

explanations can take two different forms heatmaps highlighting important image regions and prototypes (i.e., image patches) there has been relatively little work on assessing interpretable-by-design models (Kim et al., 2022)

Combining all available human studies for visual explanations we set up a questionnair: 1. Shen and Huang [62] ask users to se- lect incorrectly predicted labels with or without showing explanations 2. Nguyen et al. [49] ask users to decide whether model predictions are correct based on explanations; 3. Fel et al. [23] ask users to predict model outputs in a concurrent work different from 49 &62, we ask users to select the correct prediction out of multiple predictions to reduce the effect of confirmation bias and don't show class labels to prevent users from relying their prior knowledge (**?**) uses agreement tasks (how confident are you that the protoype from prediction looks like prototype learned during training) -¿ this looks like that -¿ However, it doesn't measure the utility of explanations in distinguishing correct and incorrect predictions, a crucial functionality of explanations in AI-assisted decision making distinction task: which class do you think the model predicts

remove the effect of human prior knowledge in our evaluations -¿ task are so fine grained / used medical trainees (Kim et al., 2022) found that participants generally believe model predictions are correct when given explanations for them

Cross-method comparison: evaluating all methods on a common task.

Study Design Introduction: Introduce the concept of prototypes + why we do this study in simple terms without technical keywords to make it understandable to layperson or persons with limited ML knowledge request optional demographic data regarding gender identity, race and ethnicity about the participant's experience with machine learning; however, no personally identifiable information was collected Objective evaluation tasks: distinction task Subjective evaluation questions: Ask if they have any complaints about how the explanations are presented Which model do you prefere in it's explanation / presentation maybe: we ask the participant to self-rate their level of understanding of the evaluated method before and after completing the task, to investigate if the participant's self-rated level of understanding undergoes any changes during the task.

Experimental design **Data & Models Human Studies** **Statistical Analysis** report the mean task accuracy and standard deviation of the participants' performance which captures the variability between individual participants' performance compare the study result to random chance and compute the p-value from a 1-sample t-test.1 When comparing results between two groups, we compute the p-value from a 2-sample t-test. Results are deemed statistically significant under p ¡ 0.05 conditions.

4. Results

4.1. Performance Evaluation

- Present the performance results of different ProtoPNet architectures on the medical datasets. - Compare these results to the baselines.

4.2. Prototype Analysis

- Analyze how the number of prototypes per class affects performance (ACC, AUC, training time). - Provide visualizations of prototypes and their matched image patches.

4.3. Expert Feedback

- Summarize the feedback from medical experts on the usability of prototypes. - Discuss the potential use cases of prototypes in clinical practice.

5. Discussion

- Interpret the experimental results. – qualitative experiments: When given multiple model predictions and explanations, participants strug- gle to distinguish between correct and incorrect predictions based on the explanations (e.g., achieving only 40% accuracy on a multiple-choice task with four options). This result suggests that interpretability methods need to be improved to be reliably useful for AI-assisted decision making. (Kim et al., 2022)

Hoffmann et al. (Hoffmann et al., 2021) highlight that prototype similarity of ProtoPNet does not correspond to semantic similarity and that this disconnect can be exploited.

we did not go through multiple iterations of UI design

- Discuss the implications of prototype-based interpretability in medical AI. - Address the limitations of the current study and propose future work. When provided explanations, participants tend to believe that the model predictions are correct, revealing an issue of confirmation bias. (Kim et al., 2022)

6. Conclusion

limitations of our work: First, we use a relatively small sample setup is still far from real-world uses of interpretability methods - Recap the research question and main findings. - Highlight the contributions of the study to the field of interpretable AI and medical imaging. - Suggest practical recommendations for implementing ProtoPNet in clinical settings.

Acknowledgements

Impact Statement

References

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., and Nensa, F. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, pp. 110786, 2023.

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

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

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., and Nushi, B. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1362–1374, 2022.

Grote, T. and Berens, P. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

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

Gupta, J. and Seeja, K. A comparative study and systematic analysis of xai models and their applications in healthcare. *Archives of Computational Methods in Engineering*, pp. 1–26, 2024.

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

Hoffmann, A., Fanconi, C., Rade, R., and Kohler, J. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*, 2021.

Kim, B., Rudin, C., and Shah, J. A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems*, 27, 2014.

Kim, S. S., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pp. 280–298. Springer, 2022.

Krupinski, E. A. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72 (5):1205–1217, 2010.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pp. 5491–5500. PMLR, 2020.

Kushniruk, A. W. Analysis of complex decision-making processes in health care: cognitive approaches to health informatics. *Journal of biomedical informatics*, 34(5): 365–376, 2001.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., and Doshi-Velez, F. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684, 2016.

Melis, D. A., Kaur, H., Daumé III, H., Wallach, H., and Vaughan, J. W. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pp. 35–47, 2021.

Paul, D., Chowdhury, A., Xiong, X., Chang, F.-J., Carlyn, D., Stevens, S., Provost, K., Karpatne, A., Carstens, B., Rubenstein, D., et al. A simple interpretable transformer for fine-grained image classification and analysis. *arXiv preprint arXiv:2311.04157*, 2023.

Potter, M. C., Wyble, B., Hagmann, C. E., and McCourt, E. S. Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76:270–279, 2014.

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

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

Simon, H. A. Rational decision making in business organizations. *The American economic review*, 69(4):493–513, 1979.

Subramanian, G. H., Nosek, J., Raghunathan, S. P., and Kanitkar, S. S. A comparison of the decision table and tree. *Communications of the ACM*, 35(1):89–94, 1992.

Wu, J., Liu, G., Wang, J., Zuo, Y., Bu, H., and Lin, H. Data intelligence: Trends and challenges. *Syst. Eng.-Theory Pract*, 40:2116–2149, 2020.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. 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. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.