HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240214

Explainable Interactive Machine Learning Using Prototypical Part Networks for Medical Image Analysis

Alec PARISE, Brian MAC NAMEE
Science Foundation Ireland Centre for Research Training in Machine Learning
School of Computer Science, University College Dublin, Dublin, Ireland
ORCiD ID: Alec Parise https://orcid.org/0009-0007-2303-3760, Brian Mac Namee https://orcid.org/0000-0003-2518-0274

Abstract. Medical imaging is a critical component of clinical decision-making, patient diagnosis, treatment planning, intervention, and therapy. However, due to the shortage of qualified radiologists, there is an increasing burden on healthcare practitioners, which underscores the need to develop reliable automated methods for interpreting medical images to reduce the time spent on commonplace cases and to support radiologists on more complex cases. Despite the development of novel computational techniques, automatically interpreting medical images remains challenging due to the subtlety and nuance of the patterns to be interpreted as well as the presence of noise and varying acquisition conditions. One promising solution to improve the reliability and accuracy of automated medical image analysis is interactive machine learning (IML), which integrates human expertise into the model training process. However, IML methods often lack compelling explanations to help users understand how a model is processing an image. To overcome this limitation, this study introduces a novel approach that leverages active learning (AL) to iteratively query for high-uncertainty samples while utilizing explanations from a prototypical part network to improve model classification. The proposed approach utilizes prototypical parts, which are snapshots of image sections, to determine an unlabelled image's class based on the presence of the prototypical parts. Interaction occurs during the selection of prototypes and the AL phase, where a set of decision rules is designed to consider the contributions of which combinations of prototypical parts are the most representative of the unlabeled image output by the AL. The proposed explainable interactive machine learning (XIL) framework empowers medical experts to interact with the model's training process, enabling more efficient and personalized learning through explanation and interaction.

Keywords. Human-in-the-loop (HITL), Interactive Machine Learning (IML), Active Learning (AL), Explainable Artificial intelligence (XAI), Explainable interactive Machine Learning (XIL), Prototypical Part Network.

1. Context

The field of interactive machine learning (IML) has gained significant attention in the medical field in recent years [3,4,5,6,7]. Training image-based machine learning models typically relies solely on automated processes to learn patterns and make predictions,

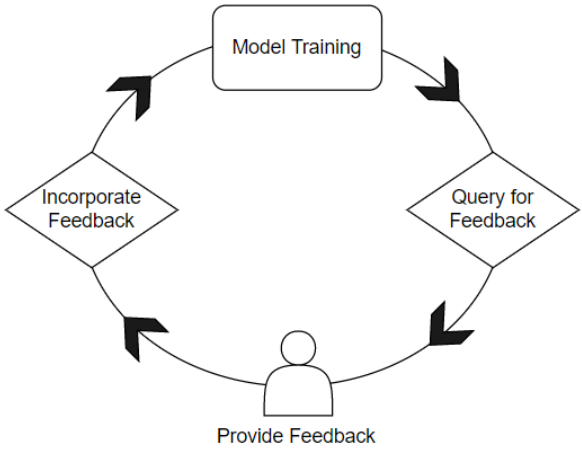

Figure 1. Basic Interactive Machine Learning (IML) Framework

offering no ability for interaction between the user and the model [7]. In contrast, IML incorporates human input and feedback into the modelling process, resulting in more interpretable and domain-specific models [3,4]. A typical IML workflow is presented in Figure 1, where the model training process is automated and periodically interspersed with interactions with a user. The user provides feedback to queries posed by the training process, which is then incorporated into another round of automated model training. Active learning [8], is a well known example of IML, and aims to reduce the number of labeled instances required to train machine learning models. However, there is an opportunity to deepen the interaction between the oracle[1] by enabling the system to accompany user feedback queries with model explanations and allowing for more sophisticated feedback than simple labels in active learning.

The imperative for explainable artificial intelligence (XAI) in medical image analysis is driven by the need for transparency and understandability in model predictions [9]. Techniques such as local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) have set benchmarks in demystifying model decisions by attributing predictions to specific features in the input data [10,11]. The integration of XAI into IML frameworks (XIL) ushers in an era where users can not only interact with models, but they can also grasp the 'why' behind model predictions. This is critical in medical diagnostics, where understanding the rationale for a particular diagnosis or treatment recommendation can significantly impact patient outcomes and clinician trust in AI tools.

Another emerging trend in XAI for medical imaging is the use of prototypes. Chen et al. [12] define a prototype as part of an image that is representative of features that capture the essential characteristics of a class. To prevent confusion with prior research that employs the term "prototype" [13], we adopt the phrase "prototypical part" to refer specifically to a portion of an image that captures essential characteristics, rather than full

[1] A human expert who actively participates in the training process of an Interactive Machine Learning (IML) model.

image prototypes. To arrive at a final classification for an unlabelled image, the evidence of multiple prototypical parts are combined.

Following this discourse, the AIMEE system [14] is one example of a family of XAI approaches that extract rules from models to help users understand how a model works, and even provide an opportunity to adapt a model. Systems like AIMEE work with numerical data only, however. There is though an opportunity to combine the use of prototypical parts with systems like AIMEE for models that work with medical images. To illustrate, we could have a rule such as "IF (prototype A) AND (prototype B) AND NOT (prototype C) THEN (Class B)". These rules could then serve to elucidate the current state of the learned model to a user, who could provide feedback to the learning process by adjusting the rules accordingly. Our work proposes to build such a system.

To complement the decision rules, we propose implementing interactive methods for training prototypical parts. The current approach outlined by Chen et al. [12] offers a relatively static methodology. However, not all prototypical parts may be readily interpretable by human specialists. For example, a model might generate a set of prototypical parts for specialists to either accept or reject. While rejected prototypical parts still hold value for the model, they may lack interpretability for human experts and consequently might not be incorporated into the Interactive Machine Learning (IML) framework. This setup would empower an oracle to exert influence in the training process by accepting or rejecting prototypical parts based on their representativeness of the class. This functionality offers an opportunity for the system to leverage the expertise of the human oracle, thereby ensuring the selection of more relevant prototypical parts.

The combination of these approaches AL, prototypical parts, and rule-based explanations—presents a holistic strategy for enhancing machine learning applications in medical image analysis. By prioritizing interpretability, efficiency, and expert integration, the framework being built in our work addresses key challenges in the field, offering a path toward models that are not only technically proficient but also clinically valuable.

2. Research Questions

This research will address the following research questions:

RQ1: When using AL for medical image classification problems, which combination of model (low, medium or high capacity), data representation (raw images or bottleneck features from a pre-trained model), and selection strategy (random, margin or Least-Confidence) leads to the most accurate models with the fewest labelled images?

When choosing an AL framework, several key factors come into play. Firstly, the nature of the dataset is crucial, as different AL strategies excel in different data contexts. For instance, uncertainty sampling may shine in image classification datasets, while ensemble-based AL could be more effective for text data. Additionally, the size of the dataset matters; some AL methods may require a larger initial labeled dataset to be effective, while others work well with smaller sets. Model capacity is also a consideration; high-capacity models may need more labeled data and time to converge, whereas simpler models might suffice with less data. Finally, the choice of sampling and query strategies—such as uncertainty sampling or query-by-committee—can significantly impact AL effectiveness.

In terms of classifier effectiveness within AL scenarios, it's hypothesized that models capable of learning from few labeled examples are preferable. Pre-trained models like ResNet50 are anticipated to offer robustness but may be computationally intensive. To address this, an alternative AL framework is proposed, leveraging bottleneck features from ResNet50 combined with a Random Forest classifier. Additionally, shallow Convolutional Neural Networks (CNNs) are seen as promising due to their flexibility in architecture and training. This study aims to validate these hypotheses by analyzing various AL frameworks and classifiers, offering insights into optimal combinations for different dataset characteristics.

RQ2: How can interpretable decision rules based on prototypical parts (rather than feature values) be created to improve model interpretability?

In order to generate informative explanations, an XIL system will necessitate interpretable explanations, such as decision rules or rankings of feature importance that are understandable by humans.the process begins by translating the prototypical parts into a vector space. This enables a direct mapping of decision rules to specific prototypes or their combinations. The mapping would guide the network in emphasizing certain prototypes over others based on the decision rules. For example, if a decision rule indicates a particular feature is highly indicative of a class, the prototypes corresponding to that feature could be weighted more heavily in the classification process.

Additionally, the feedback mechanism from AIMEE, where users can edit or propose new rules, could be used to refine the set of prototypes uses. If a user identifies a prototype that does not contribute effectively to classification or misses a critical aspect, the network could be adjusted to incorporate this feedback, either by modifying the existing rule or learning new ones that better capture the user-defined rules.

RQ3: How can decision rules be modified by users based on the presence or absence of prototypical parts and incorporated as user feedback into an Explainable interactive Machine Learning (XIL) framework?

In addressing this question, the focus is on enabling user engagement with decision rules in medical image classification within the XIL framework. Challenges include simplifying complex rules without losing effectiveness, designing intuitive interfaces, establishing effective feedback mechanisms, ensuring model interpretability, and techniques to incorporate revised rules back into a model. By prioritizing user interaction and overcoming these challenges, we aim to enhance the transparency and interpretability of machine learning models in medical image classification.

RQ4: To what extent can the integration of IML enable the discovery of more human interpretable prototypes?

In the initial pool of labeled data, prototypes are established to harness domain knowledge for the model. However, these prototypes, while valuable, may not always be interpretable to the user. To bridge this gap, a prototype ranking system is introduced, prioritizing user interpretability. Prototypes less clear to users remain important to the model's functionality but are sidelined during user interactions. By enabling user interaction with the prototypes, allowing them to rank these based on interpretability, the model ensures that users are presented with prototypes that are meaningful and understandable. This strategy aims to balance domain significance with user interpretability, ensuring that engagement with decision rules always yields interpretable prototypical parts.

Table 1. The summary of results, measured by the Area Under Learning Curve (AULC), shows the best performing approach highlighted in bold

Representation	Model	Sampling Strategy	PnemoniaMNIST	BloodMNIST	DermaMNIST	OrganMNIST3D	FractureMNIST3D
Bottleneck Features	Random Forrest	Random	0.8270	0.6399	0.6649	0.6529	0.4351
		Margin	0.8295	0.7144	0.6783	0.6456	0.4109
		Least-Confidence	0.8284	0.6341	0.6789	0.6531	0.4273
Raw Image	ResNet50	Random	0.7960	0.9194	0.6892	0.9027	0.4329
		Margin	**0.8614**	**0.9302**	0.6919	**0.9211**	0.4283
		Least-Confidence	0.8452	0.9262	**0.7134**	0.9180	**0.4542**
Raw Image	Shallow CNN	Random	0.8213	0.7421	0.6562	0.7121	0.4098
		Margin	0.7879	0.6571	0.6587	0.7558	0.3922
		Least-Confidence	0.8333	0.7488	0.6602	0.7307	0.4073

3. Methodology

Early work has addressed RQ1 and designed an experiment aimed to assess the performance of pool-based active learning using various combinations of query strategies (random, margin, and Least-Confidence), model representations (raw image and bottleneck features), and model types (random forest, 5-layer CNN, and ResNet50). The raw image representations were resized to 224x224 and were used as input for the medium and high capacity models. To ensure a balanced representation across all classes, the AL workflow began by selecting an initial subset of labeled images consisting of 20 samples in all studies. Instead of relying on human agents for labeling, a synthetic approach was used to simulate adding a labels to images per iteration. During each of 240 iterations, the four most informative unlabeled instances were labeled and added to the dataset based on the chosen query strategy. The test set, which had already been split by the authors of MedMNIST [15], was used to evaluate the model's generalization on unseen data at each iteration. Performance evaluation was based on two metrics: the area under the learning curve (AULC) and accuracy (ACC) after 100 iterations. The process was repeated for each combination of query strategy, model representation, and model type

4. Results

This experiment aimed to determine the most effective combination of model representation, model capacity, and query strategies for active learning (AL) scenarios involving medical images. Our experiment involved two image representations (bottleneck feature and raw image representations) and three model architectures (Random Forest (low capacity), a 5-layer CNN (medium capacity), and a ResNet50 (high capacity)). We employed three query strategies, namely Random, Margin, and Least-Confidence, to identify the most informative data points for labeling.

The results, summarized in Tables 1 and 2, reveal that the high-capacity ResNet50 model using raw image representation, coupled with either the margin or least-confidence query strategies, consistently achieved superior performance compared to other combinations.

In summary, our study highlights the effectiveness of employing ResNet50 with raw image representations in AL scenarios. This approach achieves impressive accuracy while requiring significantly fewer labeled samples compared to benchmark models. Furthermore, our findings underscore the importance of selecting the appropriate query

Table 2. The summary of results, measured by the Accuracy metric (ACC %), shows the best performing approach highlighted in bold.

Representation	Model	Sampling Strategy	PnemoniaMNIST	BloodMNIST	DermaMNIST	OrganMNIST3D	FractureMNIST3D
Bottleneck Features	Random Forrest	Random	83.02	62.76	65.97	68.03	40.83
		Margin	84.13	73.66	68.07	68.52	39.58
		Least-Confidence	83.33	61.36	67.98	68.53	42.08
Raw Image	ResNet50	Random	79.81	93.74	69.02	87.51	43.75
		Margin	**87.82**	**96.66**	72.15	**93.12**	40.41
		Least-Confidence	86.70	96.14	**72.76**	92.89	**45.00**
Raw Image	Shallow CNN	Random	83.81	77.05	63.48	75.78	37.08
		Margin	80.81	67.52	63.91	81.76	40.83
		Least-Confidence	86.86	79.63	64.66	82.13	40.41
Benchmark	ResNet50		85.70	95.60	73.1	85.70	49.40
	Shallow CNN		83.20	79.42	68.54	81.93	40.12

strategy for optimal AL performance. The preferred strategy may vary depending on the image types and regions, as evidenced by our results. Moving forward, we plan to incorporate both Least-Confidence and Margin-based query strategies in future studies to ensure robustness and generalizability across different datasets and to investigate the use of different pre-trained model architectures, especially those targeted at medical images.

5. Future Work

The future work and completion plan outlined in this section focuses on addressing three key research questions related to improving model interpretability and incorporating human feedback in the context of medical image analysis. RQ2 aims to explore the creation of interpretable decision rules based on prototypical parts, rather than feature values, to enhance model interpretability. The plan involves integrating the ProtoPNET algorithm into the IML framework and adapting the AIMEE framework to generate decision rules based on prototypes. This approach aims to define clear conditions for prototypical parts in unlabeled images, thereby improving the interpretability of the model. For RQ3, the objective is to investigate how decision rules can be modified by users based on the presence or absence of prototypical parts and incorporated as user feedback into an XIL framework. The plan involves exploring how oracles can modify decision rules based on prototypes to build more trustworthy models using domain expertise. This approach aims to refine the model's output to increase the expert's understanding and establish trust during the decision-making process. Finally, RQ4 seeks to assess the extent to which the integration of IML can enable the discovery of more human-interpretable prototypes. This involves incorporating IML approaches into prototype discovery algorithms to involve human experts during the generation of prototypical parts. The aim is to identify more human-interpretable prototypes, thereby enhancing the applicability, trust, and interpretability of the models.

6. Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis. 2021;71:102062.
[2] The Royal College of Radiologists. RCR Clinical Radiology Census Report 2021. Accessed on February 27, 2023. https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021.
[3] Jiang L, Liu S, Chen. Recent research advances on interactive machine learning. Journal of Visualization. 2019;22:401-417.
[4] Amershi S, Cakmak M, Knox WB, Kulesza T. Power to the people: The role of humans in interactive machine learning. Ai Magazine. 2014;35(4):105-120.
[5] Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop?. Brain Informatics. 2016;3(2):119-131.
[6] Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, Schiegg M, Ales J, Beier T, Rudy M, et al. Ilastik: interactive machine learning for (bio) image analysis. Nature methods. 2019;16(12):1226-1232.
[7] Teso S, Kersting K. Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019. p. 239-245.
[8] Settles B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences; 2009.
[9] Ghai B, Liao QV, Zhang Y, Bellamy R, Mueller K. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. arXiv preprint arXiv:2001.09219. 2020.
[10] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135-1144.
[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
[12] Chen, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems. 2019;32.
[13] Li O, Liu H, Chen C, Rudin C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1).
[14] Cornec O, Nair R, Daly E, Alkan O, Wei D. AIMEE: Interactive model maintenance with rule-based surrogates. NeurIPS 2021 Competitions and Demonstrations Track. 2022;p. 288-291.
[15] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 2023;10(1):41.