

# Extending Merlin-Arthur Classifiers for Improved Interpretability

Berkant Turan<sup>1</sup>

<sup>1</sup>*Department for AI in Society, Science, and Technology, Zuse Institute Berlin, Germany*

## Abstract

In my doctoral research, I aim to address the interpretability challenges associated with deep learning by extending the Merlin-Arthur Classifier framework. This novel approach employs a pair of feature selectors, including an adversarial player, to generate informative saliency maps. My research focuses on enhancing the classifier's performance and exploring its applicability to complex datasets, including a recently established human benchmark for detecting pathologies in X-ray images. Tackling the min-max optimization challenge inherent in the Merlin-Arthur Classifier for high-dimensional data, I will explore and apply diverse stabilization strategies to bolster the framework's robustness and training stability. Finally, the goal is to expand the framework beyond pixel-level saliency maps to encompass modalities, such as text and learned feature spaces, fostering a comprehensive understanding of interpretability across various domains and data types.

## Keywords

Interactive Classification, Mutual Information, Merlin-Arthur Classifier, Interpretability

## 1. Motivation

Over the past decade, machine learning has made tremendous progress, especially with the advancement of deep learning. But despite the astonishing advances in deep learning, major concerns have been raised about Artificial Intelligence (AI) safety in view of its large-scale application [1, 2]. One of the safety-related issues concerns the lack of interpretability of deep neural networks deployed in mission-critical tasks. To address this issue, different techniques have been developed in the field of Explainable AI (XAI), including *local* and *global* methods [3, 4, 5]. A common feature of these methods is that they are often based on heuristics [6]. Although heuristic methods have had their successes, such as unmasking biases of established classifiers [7], a large body of research has highlighted the growing concern about the disadvantages of non-formal interpretability techniques [8, 9]. Additionally, several XAI methods have shown vulnerability to manipulation through the strategic design of neural networks [10, 11, 12].

These concerns emphasize the need for the development and adoption of formal and robust explanation methods in the field of AI. Approaches to interpretability, such as Mutual Information [13] or Shapley values [14], have been proposed as more rigorous alternatives to heuristic-based methods. However, these formal techniques require a faithful modeling

---


*Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal*

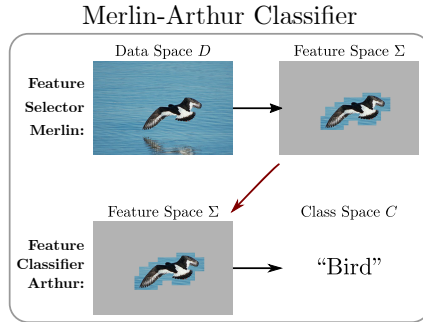
✉ turan@zib.de (B. Turan)

ORCID 0009-0001-0154-0987 (B. Turan)

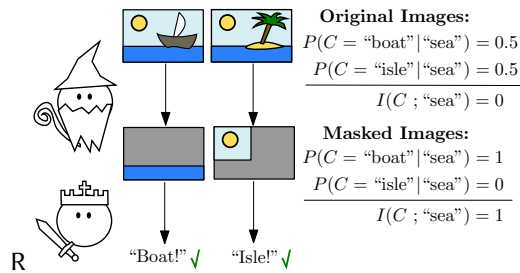


© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** The Merlin-Arthur classification system involves two interacting agents communicating via a selected feature that acts as a representation of the target class. The Illustration was taken from [6].

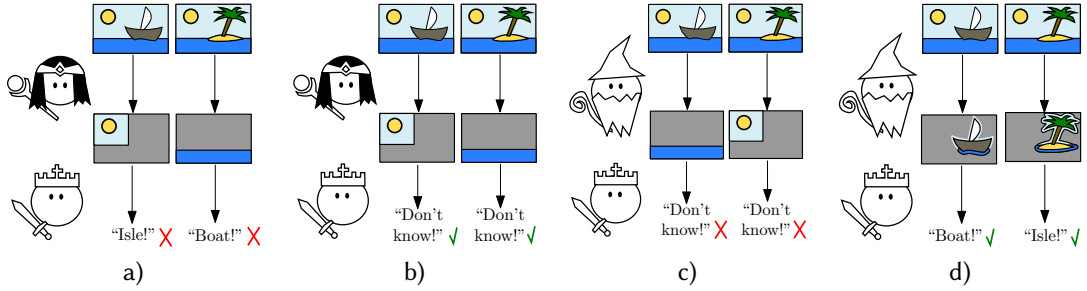


**Figure 2:** Illustration of "cheating" behavior: In the original dataset, "sea" and "sky" features appear equally in both "boat" and "island" classes. However, in Merlin's modified images, "sea" is only visible in "boat" images and "sky" in "island" images. This strong correlation enables Merlin to classify images using uninformative features, contrary to the idea of an interpretable classifier. Illustration taken from [6].

of the underlying distribution, which is often difficult for non-synthetic data. This has been practically achieved with generative models [13, 15], but there is still a requirement for trust in the underlying generative model. We can circumvent trusting a generative model by an interactive classification setup that allows us to provide bounds on the precision, which in turn provide bounds on the mutual information.

## 2. Merlin-Arthur Classifier

In this section, we discuss the *Merlin-Arthur Classifiers*, a novel classification framework developed by our research group that utilizes multi-agent interaction to allow for theoretical interpretability guarantees [6]. This framework is inspired by the Interactive Proof System (IPS), specifically the Merlin-Arthur protocol. The prover (Merlin) selects a feature from a data point and presents it to a verifier (Arthur) who determines its class, as illustrated in Figure 1. This interactive approach has already been studied in [16]. However, it has been noted that Merlin and Arthur can cooperate to achieve high accuracy with uninformative features [17, 6], see Figure 2. Thus, the adversarial aspect of interactive proof system is crucial. We introduce a second, adversarial prover (Morgana) with the objective to convince Arthur of the wrong class.



**Figure 3:** Evolution of the feature selection strategy for binary classification of “boats” and “isles”. a) Morgana exploits Arthur’s expectations to trick him using “sky” and “sea” features. b) Arthur stops giving concrete classifications to avoid being fooled. c) Arthur’s uncertainty hinders cooperation with Merlin. d) Merlin adjusts by sending only unambiguous features. Illustration is taken from [6].

In essence, our theory shows that the only strategy that Merlin and Arthur can use is one that cannot be exploited by Morgana, see Figure 3 for an illustration.

## 2.1. Formal Description

Consider a dataset  $D$  with a ground truth class map  $c : D \rightarrow C$ , where  $C$  is the set of classes, and a feature space  $\Sigma \subset 2^D$ . We say a data point  $\mathbf{x} \in D$  has the feature  $\phi \in \Sigma$  if  $\mathbf{x} \in \phi$ . We define the notion of a feature selector as a map  $M : D \rightarrow \Sigma$  such that  $\forall \mathbf{x} \in D : \mathbf{x} \in M(\mathbf{x})$ . Such a feature selector can, for example, be realized by using a pixel-based saliency method and selecting the  $k$  most salient pixels. For a visual representation, refer to Figure 1. The quality of a selected feature  $\phi$  is stated in terms of the mutual information

$$I_{y \sim \mathcal{D}}(c(\mathbf{y}); \mathbf{y} \in \phi) := H_{y \sim \mathcal{D}}(c(\mathbf{y})) - H_{y \sim \mathcal{D}}(c(\mathbf{y}) | \mathbf{y} \in \phi), \quad (1)$$

where  $H_{y \sim \mathcal{D}}(c(\mathbf{y}) | \mathbf{y} \in \phi)$  is the class conditional entropy given that  $\mathbf{y}$  contains the feature  $\phi$ , and  $\mathcal{D}$  is the data distribution on  $D$ . We extend this definition to a feature selector by taking an average over the features selected from the dataset, i.e.,  $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} H_{y \sim \mathcal{D}}(c(\mathbf{y}) | \mathbf{y} \in M(\mathbf{x}))$ . This quantity should be close to zero for a high-quality feature selector. The problem is that for most datasets, measuring  $H_{y \sim \mathcal{D}}(c(\mathbf{y}) | \mathbf{y} \in \phi)$  is not feasible, particularly for high-dimensional data. This complexity, amplified by continuous features or large datasets, results from the curse of dimensionality and therefore requires approximations or alternative methods.

In our setup, we define notions of

$$\text{Completeness: } \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[A(M(\mathbf{x})) = c(\mathbf{x})] \quad \text{and} \quad (2)$$

$$\text{Soundness: } 1 - \max_{l \in C \setminus \{c(\mathbf{x})\}} \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[A(\widehat{M}(\mathbf{x})) = l], \quad (3)$$

which are estimated on a test dataset and can be used to bound the conditional entropy for the features exchanged between Merlin and Arthur, see [6].

The challenge now lies in developing a training process for this setup that effectively handles complex data while simultaneously achieving high levels of completeness and soundness.

### 3. Research Objective

The primary benefit of Merlin-Arthur Classifiers over other existing approaches is the capability to quantitatively bound the quality of the features in terms of mutual information, even under reasonable assumptions, without relying on heuristics. Calculating the mutual information, however, may not always be feasible, particularly for complex datasets, such as Street View House Numbers (SVHN) [18], CIFAR-10 [19] or ImageNet [20]. To date, the numerical assessment of the framework has been restricted to basic datasets, such as MNIST [21] and the UCI Census dataset [22], where the mutual information between the selected feature and the target class can be computed.

This leads to the following question:

*How can Merlin-Arthur Classifiers be extended to complex datasets while maintaining a strong alignment between theoretical foundations and practical implementation?*

In my doctoral research, my objective is to expand the scope of the Merlin-Arthur Classifiers and evaluate their effectiveness on more demanding datasets, demonstrating a strong alignment between theoretical foundations and practical implementation.

#### 3.1. Dimensionality Reduction through Generative Models

The development of the Merlin-Arthur Classifiers framework would involve several key steps. First, it is crucial to identify alternative measures or techniques that can be used in place of mutual information for complex datasets. A potential research direction includes generative models, such as Variational Autoencoders (VAE) [23], to represent the high-dimensional data point on a lower-dimensional manifold. More precisely, the feature selectors - Merlin and Morgana - would not select individual pixels, but instead select features derived from the latent representation that typically correspond to more generalized features and present them to the classifier. This approach would result in a two-fold reduction in complexity, as it not only decreases the dimensionality but also permits maintaining a smaller maximum size for the feature.

*Can generative models, like Variational Autoencoders, enhance Merlin-Arthur Classifiers for complex datasets while maintaining feature quality?*

#### 3.2. Improving the Stability of Training on Complex Data

As the Merlin-Arthur Classifiers framework features a min-max objective, addressing potential instabilities that may arise during the training process is a crucial aspect of enhancing its performance on complex datasets. To stabilize the training process, it is important to explore various techniques and strategies that have been successful in addressing similar issues in other models with min-max objectives.

One prominent example is Generative Adversarial Networks (GANs) [24], which have been the subject of extensive research on stabilizing min-max objectives [25]. Drawing from this research, the stabilization of the Merlin-Arthur Classifiers can be achieved by incorporating

a combination of techniques, including diverse activation functions (e.g., Leaky ReLU), using spectral normalization, and employing various optimizers or replay strategies to enhance training stability [26]. These techniques can be tailored and integrated into the Merlin-Arthur framework to ensure a robust training process.

*Can we extend Merlin-Arthur classification to complex data in a stable way?*

### 3.3. Validation and Performance Comparison

Once suitable alternatives have been identified and the framework has been adapted, validation of the adapted framework is essential and should involve diverse datasets, spanning multiple domains and complexity levels, to demonstrate the framework's robustness and versatility.

For example, a recent study in the field of medical imaging has established a human benchmark for detecting pathologies in X-ray images. This study found that all investigated state-of-the-art interpretability methods lack accuracy and reliability [27]. Therefore, it is essential to benchmark the extended Merlin-Arthur Classifiers framework against existing methods such as LIME [4], SHAP [14], and LRP [28] to showcase its potential advantages and limitations in the medical domain. By doing so, the effectiveness and reliability of the framework can be better understood and established.

*How can we rigorously validate the extended Merlin-Arthur Classifiers framework across diverse datasets? Can we compare and identify synergies between the framework and other XAI methods?*

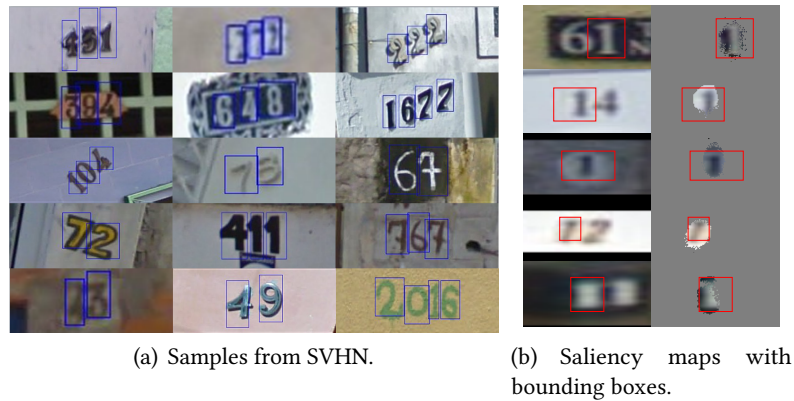
## 4. Results and Contributions

In the course of my doctoral research, I have already made progress towards achieving the objectives and key steps outlined earlier. In this section, I will present preliminary results and contributions that have been made to date.

### 4.1. Feature Quality Measurement

To assess the quality of the feature selection, appropriate datasets must be identified. While the lack of a ground truth is a common issue in XAI, we addressed this by using a modified version of the UCI Census dataset in our preprint [6]. For image data, popular datasets like CIFAR-10 and ImageNet are often used, with bounding boxes serving as the ground truth to calculate Intersection over Union (IoU) between the saliency map and the box [29].

We opted for the original SVHN dataset, which has variable-resolution color images with digits at different locations, as a better alternative to images where target objects dominate the entire image. We trained the Merlin-Arthur Classifier to distinguish images containing the digit "1" and generate saliency maps that highlight regions where the digit appears, as shown in Figure 4(b). Comparing these maps with bounding boxes, we calculated the IoU to assess map quality. Our preliminary results suggest that the Merlin-Arthur Classifiers framework effectively highlights the target regions of interest in the SVHN dataset. However, further improvements are needed for complex datasets and classification tasks.



**Figure 4:** (a) Original sample images from the full number SVHN dataset, sourced from [18]. (b) Saliency map pairs, generated using the Merlin-Arthur Classifier framework, with bounding boxes highlighting the location of the digit “1”. The binarized saliency maps effectively emphasize the regions where the digit “1” appears.

## 5. Future Directions and Potential Contributions

As we continue to extend the Merlin-Arthur Classifiers framework to handle complex datasets, there are several potential contributions that this research can make to the field of AI and XAI in particular:

1. Establishing a robust and versatile framework that can handle a wide range of datasets and classification tasks, thereby enhancing the applicability of Merlin-Arthur Classifiers in real-world scenarios.
2. Comparing the Merlin-Arthur Classifiers framework with other methods, highlighting strengths and weaknesses, and guiding future explainable AI research. Specifically, evaluating its applicability and effectiveness on medical benchmarks, such as detecting pathologies in X-ray images, where other methods have underperformed.
3. Exploring challenges, such as transcending pixel-level relevance by incorporating text-based agent conversations or leveraging Variational Autoencoder feature embeddings, to significantly advance the capabilities and impact of the Merlin-Arthur Classifiers framework.

By addressing these challenges, my doctoral research aims to advance Merlin-Arthur Classifiers, broadening their applicability, impact, and potential in XAI.

## Acknowledgments

I would like to express my gratitude to Stephan Wäldchen for his invaluable guidance in shaping this doctoral proposal. My thanks also go to Prof. Dr. Sebastian Pokutta for his diligent supervision, and to Kartikey Sharma for his constructive feedback during this initial phase.

## References

- [1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety, arXiv preprint arXiv:1606.06565 (2016).
- [2] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, *ACM Trans. Interact. Intell. Syst.* 11 (2021). URL: <https://doi.org/10.1145/3387166>. doi:10.1145/3387166.
- [3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (2018). doi:10.1145/3236009.
- [4] M. T. Ribeiro, S. Singh, C. Guestrin, "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, KDD '16*, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: <https://doi.org/10.1145/2939672.2939778>. doi:10.1145/2939672.2939778.
- [5] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., *Queue* 16 (2018) 31–57. doi:10.1145/3236386.3241340.
- [6] S. Wäldchen, K. Sharma, M. Zimmer, B. Turan, S. Pokutta, Formal interpretability with Merlin-Arthur Classifiers, arXiv preprint arXiv:2206.00759 (2023).
- [7] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, *Nature communications* 10 (2019) 1096.
- [8] J. Marques-Silva, A. Ignatiev, Delivering trustworthy AI through formal XAI, *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (2022) 12342–12350. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21499>. doi:10.1609/aaai.v36i11.21499.
- [9] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20*, Association for Computing Machinery, New York, NY, USA, 2020, p. 180–186. URL: <https://doi.org/10.1145/3375627.3375830>. doi:10.1145/3375627.3375830.
- [10] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods (2020). doi:10.1145/3375627.3375830.
- [11] B. Dimanov, U. Bhatt, M. Jamnik, A. Weller, You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods, in: *SafeAI@AAAI, 2020*.
- [12] J. Heo, S. Joo, T. Moon, Fooling neural network interpretations via adversarial model manipulation, in: *Neural Information Processing Systems*, 2019.
- [13] J. Chen, L. Song, M. J. Wainwright, M. I. Jordan, Learning to explain: An information-theoretic perspective on model interpretation, 2018. arXiv:arXiv preprint arXiv:1802.07814.
- [14] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.
- [15] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations,

- Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>. doi:10.1609/aaai.v32i1.11491.
- [16] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, arXiv preprint arXiv:1606.04155 (2016).
- [17] M. Yu, S. Chang, Y. Zhang, T. Jaakkola, Rethinking cooperative rationalization: Introspective extraction and complement control, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019. doi:10.18653/v1/D19-1420.
- [18] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. URL: <http://ufldl.stanford.edu/housenumbers>, accessed: 2022-02-23.
- [19] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, 2009.
- [20] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012).
- [21] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324.
- [22] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: <http://archive.ics.uci.edu/ml>.
- [23] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. arXiv:<http://arxiv.org/abs/1312.6114v10>.
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf).
- [25] M. Wiatrak, S. V. Albrecht, A. Nystrom, Stabilizing generative adversarial networks: A survey, arXiv preprint arXiv:1910.00927 (2019).
- [26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, arXiv preprint arXiv:1606.03498 (2016).
- [27] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, et al., Benchmarking saliency methods for chest x-ray interpretation, Nature Machine Intelligence 4 (2022) 867–878.
- [28] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) 1–46. doi:10.1371/journal.pone.0130140.
- [29] P. Dabkowski, Y. Gal, Real time image saliency for black box classifiers, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf).