CAIPI in Practice: Towards Explainable Interactive Medical Image Classification

Emanuel Slany[1(✉)], Yannik Ott[1], Stephan Scheele[1], Jan Paulus[2], and Ute Schmid[1]

[1] Fraunhofer IIS, Fraunhofer Institute for Integrated Circuits IIS, Project Group Comprehensible Artificial Intelligence, Bamberg, Germany
{emanuel.slany,yannik.ott,stephan.scheele,ute.schmid}@iis.fraunhofer.de
[2] Faculty of Electrical Engineering, Precision Engineering, Information Technology, Nuremberg Institute of Technology Georg Simon Ohm, Nuremberg, Germany
jan.paulus@th-nuernberg.de

Abstract. Would you trust physicians if they cannot explain their decisions to you? Medical diagnostics using machine learning gained enormously in importance within the last decade. However, without further enhancements many state-of-the-art machine learning methods are not suitable for medical application. The most important reasons are insufficient data set quality and the black-box behavior of machine learning algorithms such as Deep Learning models. Consequently, end-users cannot correct the model's decisions and the corresponding explanations. The latter is crucial for the trustworthiness of machine learning in the medical domain. The research field explainable interactive machine learning searches for methods that address both shortcomings. This paper extends the explainable and interactive CAIPI algorithm and provides an interface to simplify human-in-the-loop approaches for image classification. The interface enables the end-user (1) to investigate and (2) to correct the model's prediction and explanation, and (3) to influence the data set quality. After CAIPI optimization with only a single counterexample per iteration, the model achieves an accuracy of 97.48% on the Medical MNIST and 95.02% on the Fashion MNIST. This accuracy is approximately equal to state-of-the-art Deep Learning optimization procedures. Besides, CAIPI reduces the labeling effort by approximately 80%.

Keywords: XAI · Interactive learning · CAIPI · Image classification

1 Introduction

Medical diagnostics based on machine learning (ML), such as Deep Learning (DL) for visual cancer detection, have become increasingly important in the last decade [7]. However, on the one hand, clinicians are rarely experts in implementing ML models, on the other hand, even if the data sets are of high quality, they

are rarely intuitive for ML engineers. Additionally, many state-of-the-art DL algorithms are black-boxes to end-users. For ML in critical application domains like medical diagnostics, it is crucial to close the gap between clinical domain expertise and engineering-heavy ML methods.

This paper aims to enable domain experts such as clinicians to train and apply trustworthy ML models. Based on the ML pipeline, starting from data preparation to the decision-making process, we formulate three core requirements for the domain of medical diagnostics: 1) First, it is important to keep the quality of the data under control. 2) Secondly, it is necessary that the decisions of the ML model are disclosed in a transparent way - known as explainable machine learning (XAI) - where it is crucial in critical applications that models make the right decisions for the right reasons. 3) Finally, the clinical expert should be able to be involved in the optimization process [3] to allow *interactive* correction of both, explanations and decisions. Such interactive ML methods are closely related to active learning [15], in which instance and label selection occur in the interaction between algorithm and agent. By meeting these requirements, the user gains end-to-end control over the entire ML process, involving the clinical expert *interactively* in the ML process (human-in-the-loop).

We aim to combine both, explainable and interactive machine learning, denoted by *eXplainable Interactive Machine Learning* (XIML) [16]. Our domain of interest is the classification of medical images from diagnostics in everyday clinical practices, such as classifying computer tomography scans into their corresponding categories like e.g. abdomen, chest, brain, etc. For our use case, we are interested in providing a visual explanation for such a categorical classification and furthermore allow the user to correct both, the classification and the explanation.

Our core contribution focuses on four research questions about improving the applicability and efficiency of the CAIPI [16] algorithm, exemplary applied to the domain of medical diagnostics. CAIPI enables model optimization to be performed while interactively including user feedback using generated counterexamples for predictions and explanations:

(R1) Do explanation corrections enhance the model's performance [16]?
(R2) Do explanation corrections improve the explanation quality?
(R3) Does CAIPI benefit from explanation corrections for wrong predictions?
(R4) Is CAIPI beneficial compared to default DL optimization techniques?

We will outline the benefits of XIML for the medical domain in Sect. 2, recap the most important of CAIPI's core concepts and our extensions in Sect. 3, and describe our experimental setting in Sect. 4. We will answer the research questions in Sect. 5. Section 6 will discuss the results. Finally, Sect. 7 will conclude our work and summarize future research questions.

2   Related Work

Human-in-the-loop approaches provide benefits for the application of ML in the medical domain. Even if ML systems for medical diagnostics account for

expert knowledge, they can still suffer from a lack of trust, since knowledge bases can be manipulated [5]. The authors [5] propose an architecture that allows an authorized user to enrich the knowledge base while protecting the system from manipulation. Although we do not provide an explicit architecture, our scope is closely related, as we aim to enable experts to control the data quality, and to monitor and correct the behavior of the ML system. Apart from trustworthiness, another major benefit of interactive ML algorithms lies in their efficiency. For instance, extracting patient groups is more efficient when using sub-clustering with human expert knowledge compared to traditional clustering [4].

A central explanatory method, which CAIPI is based on, is called *Local Interactive Model-agnostic Explanation* (LIME) [10], which samples local interpretable features to fit a simplified and explainable surrogate model where the surrogate model's parameters become human interpretable. Although LIME is one of the most famous local explanation methods, there are alternatives such as the *Model Agnostic suPervised Local Explanations* (MAPLE) method [9] that relies on linear approximation of Random Forest models for explanations.

It is worth noting that local explanation procedures are limited to explain single prediction instances only. Also, they require additional explanatory models, which also introduce uncertainty [11]. The models do not explain the black-boxes per se, since they are ML models with different optimization objective for themselves. In contrast to local explanations, global explanations aim to explain the prediction model in general. This can be achieved by approximating the complex black-box model by a simpler interpretable model. An algorithm to approximate complex models with decision trees is proposed by [1]. The major benefit of global explanations is that the interpretable model mimics the explicit complex model. However, even if the resulting models are simpler, their interpretation still requires basic ML knowledge, which apart from the computational complexity is the major drawback for global explanations.

The authors of [14] also extend the CAIPI algorithm specifically for DL use cases. They introduce a loss function with additional regularization term. Large gradients in regions with irrelevant features are penalized. Explanations for DL models for medical image classification can also be generated with inductive logic programming [13]. The connection of this paper with both of the previous papers appears to be interesting for future research.

3 Practical, Explainable, and Interactive Image Classification with CAIPI

In this section, we first recapitulate the mathematical foundations of LIME and the operation of the CAIPI algorithm. Secondly, we will discuss our extension of the CAIPI algorithm for application to complex and large image data.

The extensions will be derived by solving two problems that frequently occur with CAIPI in practice: First, default CAIPI only receives explanation corrections, if the prediction is correct but the explanation is wrong [16]. This seems to be inefficient, as wrong predictions are also made during the optimization.

Therefore, we extend CAIPI such that users can also correct explanations if the prediction is wrong. Secondly, a major contribution of our work lies in the simplification of the human-algorithm-interaction. In practice, optimization procedures are too complex for domain experts such as physicians when they depend on human interaction. To overcome this issue, we provide an universally applicable user interface.

LIME [10] exploits an interpretable surrogate model to construct explanations for predictions of a complex model. The representation of an instance is defined by $x \in \mathbb{R}^d$. The features of an instance are transformed into an interpretable representation $x' \in \{0, 1\}^d$, where 0 indicates the absence and 1 the presence of a super-pixel. Super-pixels are contiguous patches of similar pixels in an image. Correspondingly, z is a sample generated around the original representation and z' a sample around the interpretable representation. The term $\pi_x(z)$ is a proximity measure between x and z. The complex model is denoted by $f(x)$ and the local explanatory model by $g \in G$, respectively, where G represents the aggregation of local explanatory models for $f(x)$. The term $\Omega(g)$ penalizes increasing complexity of the explanatory model. The objective function of LIME in (1) aims to minimize the sum of the loss function and the penalty.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

The locality-aware loss is defined in (2). Locality-awareness means to account for the sampling region around the representation. This is ensured by $\pi_x(z)$, which is calculated by an exponential normalized distance function.

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \tag{2}$$

We make use of the Quick Shift algorithm [17] to partition an input image into super-pixels and the Sparse-Linear Approximation algorithm [10] together with the loss function (2) to generate explanations.

CAIPI [16] distinguishes between a labeled data set L and an unlabeled data set U. It uses four components: 1) The *Fit* component trains a model with L. 2) *SelectQuery* selects a single instance from U. Typically, the label belonging to this instance maximizes the loss reduction for the next optimization step. For that, we predict the instances of U and choose the instance with the lowest prediction score. 3) *Explain* applies the LIME algorithm and shows the prediction with its corresponding explanation. 4) Depending on the user input, the *ToCounterExamples* component generates counterexamples. The selected instance is removed from U and added to L together with the generated counterexamples.

We propose an image-specific data augmentation procedure. Figure 1 shows decisive features of a computer tomography scan of the chest. We scale, rotate, and translate the decisive features. The order of the augmentation is fixed. Their parameters are random with the constraint that the resulting image must fit

(a) Original.   (b) Features.   (c) Scaling.   (d) Rotating.   (e) Translation.
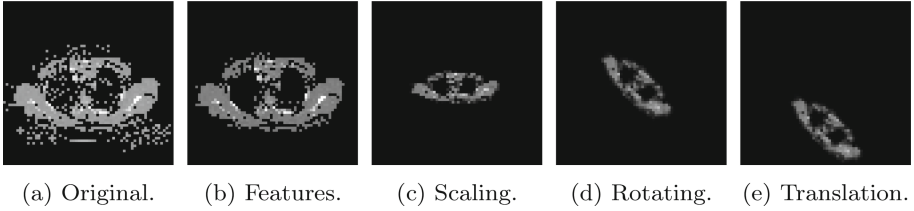
Fig. 1. Data augmentation for counterexamples. Relevant features (b) are extracted from the original image (a). The features are scaled (c), rotated (d) and translated (e).

Algorithm 1. CAIPI algorithm [16]

Require: labeled examples L, unlabeled examples U, iteration budget T
1: $f \leftarrow Fit(L)$
2: **repeat**
3:     $x \leftarrow SelectQuery(f, U)$
4:     $\hat{y} \leftarrow f(x)$
5:     $\hat{z} \leftarrow Explain(f, x, \hat{y})$
6:     Present $x, \hat{y}$ and $\hat{z}$ to the user
7:     Obtain y and explanation correction C
8:     $\{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{c} \leftarrow ToCounterExamples(C)$
9:     $L \leftarrow L \cup \{(x, y)\} \cup \{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{c}$
10:     $U \leftarrow U \backslash (\{x\} \cup \{\overline{x}_i\}_{i=1}^{c})$
11:     $f \leftarrow Fit(L)$
12: **until** max. number of iterations T or min. quality of f is reached
13: **return** f

---

completely into the original frame. The augmentation is performed with Albumentations [2]. Within CAIPI, this procedure is applied to the decisive features once when the image from U is appended to L.

In the CAIPI optimization process in Algorithm 1, the user provides feedback to the most informative instance in each iteration. The model is then retrained with the additional information. The procedure terminates when reaching a certain prediction quality of f or the maximum number of iterations.

CAIPI distinguishes between three prediction outcome states: right for the right reasons (RRR), right for the wrong reasons (RWR), and wrong (W) [16]. Whereas RRR does not require additional user input, CAIPI asks the user to correct the label for W, and to correct the explanation for RWR. RWR results in augmented counterexamples, which only contain the decisive features.

At this point, we propose the following adjustment: We require the user to provide the correct label as well as the correct explanation for case W. Theoretically, this adjustment makes the optimization process more efficient, as counterexamples are generated in each iteration if either the label or the explanation or both are wrong.

Figure 2 illustrates our proposed user interface. The depicted example image is a computer tomography scan of the chest and is displayed together with its prediction (Fig. 2a). Button *Explanation* displays the LIME result as shown in Fig. 2b. The user can then choose whether the image was predicted correctly or not (buttons *True* or *False(W)*, respectively). In case of a correct prediction, we further distinguish between right (*True(RR)*) and wrong (*True(WR)*) reasons. This distinction maps exactly to the three cases RRR, RWR and W from CAIPI.

Figure 2a shows that the image was predicted correctly. However, as Fig. 2b indicates, the explanation is at least partly wrong, i.e., the instance can be considered as RWR. The corresponding button *True(WR)* opens the annotation mode (Fig. 2c), where the user can correct the explanation. Afterwards, a newly generated explanation can be evaluated as depicted in Fig. 2d. Confirming a correction starts CAIPI's *ToCounterExamples* method, which is in our case the proposed data augmentation procedure (Fig. 1). Note, that the same interaction applies to the W case, where the interface additionally asks for the correct label. For RRR, contrary, the correction mode (Fig. 2d and 2d) is concealed from the user. The remaining procedure is constant with the slight modification that no counterexamples are generated.

The extension we propose offers great benefits for CAIPI. First of all, CAIPI can be operated by end-users. Secondly, it fulfills all essential requirements defined in the Introduction, Sect. 1. CAIPI shows its prediction and explanation to the end-user in each optimization iteration, and if necessary the end-user can correct both. Furthermore, the end-user (which typically is a domain expert) is directly responsible for the data set quality, as CAIPI asks to add instances to the training data set iteratively and the end-user can ensure correct labels and emphasize correct explanations.

4   Experiments

For our experiments, we use two classes of the Medical MNIST data set [8,18] (chest and abdomen computer tomography scans) and two classes of the Fashion MNIST data set [19] (pullover and T-shirt/top). By this selection, we want to emphasize a challenging binary classification task. The extension to categorical data is left as future work due to simplicity during the evaluation process.

We use a fairly simple convolutional neural network (CNN) as DL model in all experiments. It has a single convolutional layer with only 2 filters, a 9x9 kernel size, and stride parameter 1. It is followed by a pooling layer with kernel-size 8x8 and stride parameter 8. It follows a single linear layer with 98 neurons, a dropout layer with 0.5 dropout-rate, and two fully-connected layers with 16 and 2 neurons. All training procedures use batch size 64 and 5 epochs. We use a binary cross entropy loss function, the Adam optimization algorithm [6], and a learning rate of 0.001. The CNN corresponds to the function f in CAIPI.
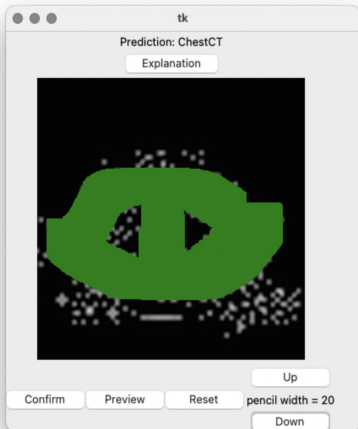
Each CAIPI optimization starts with 100 preliminary labeled instances L_0. The maximum number of iterations is set to 100. We do not specify any other stop criterion. R1 until R3 are evaluated with an alternating number of counterexamples c. The number of counterexamples per iteration is $c = \{0, 1, 3, 5\}$.
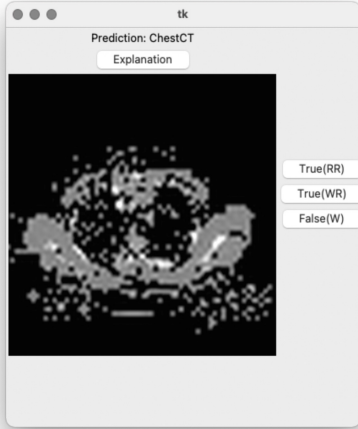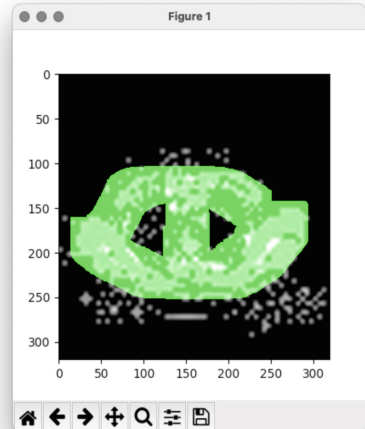
(a) Prediction of the classifier.
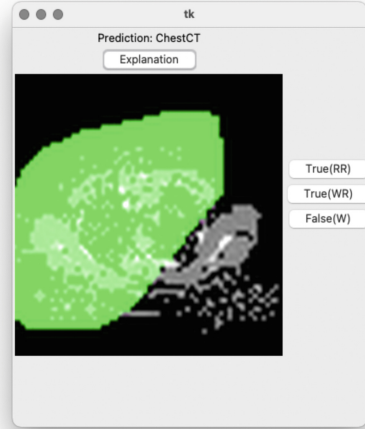
(b) LIME explanation.

(c) Correction.

(d) Corrected instance.

Fig. 2. User interface. The prediction (a) and the explanation (b) is presented to the user. The user can correct the model's prediction and explanation in the annotation mode (c). The corrected instance can be displayed (d).

We show results on a domain-related data set, the Medical MNIST, and on the Fashion MNIST, a well-known benchmark data set. We ensure balanced classes in both data sets. Since CAIPI has 100 iterations, and L_0 has 100 instances,

Table 1. Maximum accuracy (%) by number of counterexamples conditioned on data sets and modes. The mode RWR only generates counterexamples for Right predictions with Wrong Reasons, whereas the RWR + W mode generates counterexamples additionally for Wrong predictions.

Mode	Data	Counterexamples c			
		0	1	3	5
RWR	Medical MNIST	96.02	95.42	94.51	96.75
RWR + W	Medical MNIST	96.83	97.48	96.92	**97.52**
RWR	Fashion MNIST	**95.64**	94.79	95.40	94.95
RWR + W	Fashion MNIST	94.33	95.02	94.24	94.10

the final training data set will contain 200 different instances. Depending on the user input, the training data set size will increase due to counterexamples.

We evaluate the prediction quality of our model in each optimization iteration with CAIPI with the accuracy metric on dedicated test data sets with size 6,000 for the Medical MNIST and size 4,200 for the Fashion MNIST. We also created test data sets to evaluate the explanation quality. Here, both test data sets have size 200. We annotated the true explanation for all instances. For the evaluation of the explanation quality, we use the *Intersection over Unions* (IoU) metric. IoU lies in the interval $[0, 1]$, where 0 is a completely incorrect and 1 a perfect explanation. This means we divide the intersection of the LIME explanation and the ground truth explanation by their union. We consider the average non-zero explanation score. Non-zero stands for excluding false predictions, since we cannot assume correct explanations for false predictions, and average means dividing by the number of instances with non-zero explanation score. We compare the prediction and explanation ability for generating counterexamples only for RWR predictions versus generating counterexamples for predictions that are either RWR or W.

Furthermore, we conduct a benchmark test by using the identical DL setting and train a model with 14,000 training instances for the Medical MNIST and evaluated on 6,000 test samples. Correspondingly, we trained with 9,800 instances from the Fashion MNIST and tested with 4,200 instances.

5   Results

Table 1 clearly shows that the prediction quality of our model does not benefit from an increasing number of counterexamples, as the maximum accuracy is approximately stable over the runs. Also, Table 2 shows no clear trend towards an increasing explanation quality for greater numbers of counterexamples. Thus, R1 and R2 can be negated based on the experimental setting in Sect. 4. Similarly, for R3, the adjustment of providing explanation corrections also for wrong predictions does not have positive impact on either the explanation nor the prediction quality.

Table 2. Maximum average non-zero explanation score (%) by number of counterexamples conditioned on data sets and modes. The mode RWR only generates counterexamples for Right predictions with Wrong Reasons, whereas the RWR + W mode generates counterexamples additionally for Wrong predictions.

Mode	Data	Counterexamples c			
		0	1	3	5
RWR	Medical MNIST	41.59	**44.74**	40.36	42.87
RWR + W	Medical MNIST	39.64	42.37	41.12	40.47
RWR	Fashion MNIST	65.24	64.41	64.40	65.48
RWR + W	Fashion MNIST	64.43	65.76	63.10	**66.38**

For state-of-the-art DL optimization, we achieve an accuracy of 94.67% for the Medical MNIST and 95.26% for the Fashion MNIST. This accuracy is approximately equal to the CAIPI results in Table 1. Besides, CAIPI requires significantly less training data (200, plus counterexamples) than traditional optimization $(14,000,$ respectively $9,800)$. With respect to R4, this is clear evidence that CAIPI influences the optimization process positively.

6   Discussion

R1 and R2 show no clear trend in favor of an increasing number of counterexamples, despite [16] states otherwise. Teso and Kersting [16] induce decoy pixels with colors corresponding to the different classes into their training data set and they randomize the pixel colors in the test set. Whenever they create counterexamples, they also randomize the decoy pixel color. We investigate R1 and R2 without prior data set modification. Table 1 shows that default active learning $(c = 0)$ positively influences the learning behavior to such an extent that there is hardly space for improvement by extending it to XIML. This means that for future evaluations the use cases must be sufficiently complex so that default active learning does not provide satisfactory results.

We evaluate R2 by a dedicated test set containing annotated true explanations. We use IoU to estimate the quality of the explanation in percent by dividing the area of intersection by the area of union. The idea is, if predicted and annotated explanations are congruent to each other, the explanation is perfect. We frequently observe perfectly negative explanations, meaning that the predicted explanation highlights all image parts apart from the annotated explanation. From a human perspective, this explanation is perfect, for IoU it is a completely wrong explanation. Determining the quality of explanations is a prominent research field and future work should evaluate additional metrics.

For R3, Table 1 and 2 show that explanation corrections for RWR and W do not differ significantly compared to explanation corrections for only RWR. We argue that including more counterexamples (RWR + W) can be a chance to build more robust data set. The robustness from a statistical perspective was

not addressed in this paper. It must be included in future evaluations besides accounting for the priory mentioned discussion points. Similar to earlier discussion points, R4 can be re-evaluated in more complex use cases.

The data augmentation procedure plays also a major role. The procedure defined in Fig. 1 will create training data, which are fundamentally different to the test data, since both data sets, Medical MNIST and Fashion MNIST, contain relatively centralized images. The idea behind the proposed procedure is to force the model to account for the decisive features without considering the feature position. This can be enhanced by random transformations in every epoch compared to a single transformation when the counterexamples are generated. We also expect improvement by including further constraints to make the resulting counterexamples more realistically.

Finally, our main contribution is the simplification of the human-algorithm-interaction with the introduced interface. We support this on theoretical basis, as the application of CAIPI with our interface fulfills requirements, which we defined in the Introduction, Sect. 1. Furthermore, we give practical evidence via demonstration in Sect. 3. From a psychological point of view, this is insufficient. Therefore, our interface should be subject of psychological studies in future.

7 Conclusion and Future Work

We extended the CAIPI algorithm by accounting additionally for explanation corrections if the predictions are wrong. Moreover, we introduced an user interface for a human-in-the-loop approach for image classification tasks. The interface enables the end-user (1) to investigate and (2) to correct the model's prediction and explanation, and (3) to influence the data set quality.

The experiments show that the predictive performance of state-of-the-art DL methods is met, even though the required training data set size decreases. According to our findings, the correlation between an increasing amount of counterexamples and higher predictive and explanatory quality does not hold. The introduced extension that creates counterexamples also for wrong predictions can help to build more robust data sets but does not increase the predictive nor the explanatory quality. The proposed interface is a promising extension for medical image classification tasks using CAIPI. The interface appears to be transferable to every XIML approach exploiting local explanations. Evidently, CAIPI as well as the proposed interface is transferable to any other image classification task.

The most obvious improvement is the generalization to categorical image data. This appears to be a minor adjustment. It was neglected in this paper for the sake of simplicity during evaluation of the experiments. Future research should also address wrong explanations. This can be accomplished by connecting this paper with [14]. Another prominent research subject is the CAIPI algorithm for itself. As the CAIPI algorithm can be considered as feedback-reliable data augmentation procedure, it could be continuously adjusted and modified. Here, research subjects can be instance selection, local explanation, or data augmentation methods. More sophisticated methods than simple IoU are necessary to estimate the visual explanation quality more accurately.

Further adjustments can be separated into three groups. First, the interface can be evaluated in psychological studies. Second, the computational efficiency of XIML methods can be increased by connecting them with online learning algorithms such as [12]. And third, the connection of inductive logic programming like in [13] with human-in-the-loop ML procedures is a promising research area.

References

1. Bastani, O., Kim, C., Bastani, H.: Interpreting blackbox models via model extraction (2017). http://arxiv.org/abs/1705.08504

2. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020). https://doi.org/10.3390/info11020125

3. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016). https://doi.org/10.1007/s40708-016-0042-6

4. Hund, M., et al.: Analysis of patient groups and immunization results based on subspace clustering. In: Guo, Y., Friston, K., Aldo, F., Hill, S., Peng, H. (eds.) BIH 2015. LNCS (LNAI), vol. 9250, pp. 358–368. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23344-4_35

5. Kieseberg, P., Schantl, J., Frühwirt, P., Weippl, E., Holzinger, A.: Witnesses for the doctor in the loop. In: Guo, Y., Friston, K., Aldo, F., Hill, S., Peng, H. (eds.) BIH 2015. LNCS (LNAI), vol. 9250, pp. 369–378. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23344-4_36

6. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6980

7. Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., Fotiadis, D.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13** (2014). https://doi.org/10.1016/j.csbj.2014.11.005

8. Lozano, A.P.: Medical MNIST Classification (2017). https://github.com/apolanco3225/Medical-MNIST-Classification

9. Plumb, G., Molitor, D., Talwalkar, A.S.: Model agnostic supervised local explanations. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, 3–8 December 2018, pp. 2520–2529 (2018). https://proceedings.neurips.cc/paper/2018/hash/b495ce63ede0f4efc9eec62cb947c162-Abstract.html

10. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778

11. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

12. Sahoo, D., Pham, Q., Lu, J., Hoi, S.C.H.: Online deep learning: learning deep neural networks on the fly. In: Lang, J. (ed.) Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, 13–19 July 2018, pp. 2660–2666. ijcai.org (2018). https://doi.org/10.24963/ijcai.2018/369

13. Schmid, U., Finzel, B.: Mutual explanations for cooperative decision making in medicine. Künstliche Intelligenz **34**(2), 227–233 (2020). https://doi.org/10.1007/s13218-020-00633-2

14. Schramowski, P., et al.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nat. Mach. Intell. **2**(8), 476–486 (2020). https://doi.org/10.1038/s42256-020-0212-3

15. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (2012). https://doi.org/10.2200/S00429ED1V01Y201207AIM018

16. Teso, S., Kersting, K.: Explanatory interactive machine learning. In: Conitzer, V., Hadfield, G.K., Vallor, S. (eds.) Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, 27–28 January 2019, pp. 239–245. ACM (2019). https://doi.org/10.1145/3306618.3314293

17. Vedaldi, A., Soatto, S.: Quick shift and kernel methods for mode seeking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 705–718. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88693-8_52

18. Yang, J., Shi, R., Ni, B.: MedMNIST classification decathlon: a lightweight AutoML benchmark for medical image analysis. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 191–195 (2021)

19. Zalando SE: Fashion MNIST (2017). https://www.kaggle.com/zalando-research/fashionmnist