Journal of Industrial Information Integration 45 (2025) 100827

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Full Length Article

# On the disagreement problem in Human-in-the-Loop federated machine learning

Matthias Huelser [a,b], Heimo Mueller [a,b], Natalia Díaz-Rodríguez [c], Andreas Holzinger [d,a,b,*]

[a] Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Austria
[b] Information Science and Machine Learning Group, Diagnostics and Research Institute for Pathology, Medical University Graz, Austria
[c] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain
[d] Human-Centered AI Lab, Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, University of Natural Resources and Life Sciences Vienna, A-1190 Vienna, Austria

## ARTICLE INFO

*Keywords:*
Human-Centric AI
Human-in-the-Loop
Human–machine interaction
Counterfactuals
Active Learning

## ABSTRACT

The popularity of Artificial Intelligence (AI) has risen sharply in recent years, revolutionizing applications in most sectors with unprecedented functionalities. Milestones and achievements like ChatGPT demonstrate not only the impressive capabilities of AI, but also how accessible such technologies have become in recent times. However, the success of AI applications depends heavily on the underlying information integration processes. Among the most important processes are the training of the AI model at the core of the application and the collection and pre-processing of training data. In particular, the task of collecting high-quality training data can be very costly and resource-intensive, as in many cases large amounts of data have to be annotated manually. Human annotators must have extensive expertise for certain tasks in order to provide high-quality training data. In this paper, we present a framework to maximize the efficiency of human experts in a Machine Learning (ML) scenario, with the aim of optimizing the use of human expertise in active learning. This is done by constantly measuring the quality of human experts' input, as well as by involving human annotators only when needed. We showcase the benefits of our proposed framework by applying it to a problem in image classification, proving its usefulness to reduce the cost of annotating training data. The source code of the framework is publicly available at https://github.com/human-centered-ai-lab/app-HITL-annotator.

### Article summary

What is already known to the research community?

1. The importance of high quality training data in machine learning processes and the techniques to obtain such data.
2. Well known techniques on how to integrate human expert knowledge into ML processes such as active learning.
3. The difficulty of dealing with manual annotation processes, especially when it comes to handling disagreement between multiple experts, as well as finding a consensus and obtaining a ground truth.
4. Well-known techniques of explainable AI, especially the idea of counterfactuals in the area of image classification.

What this paper contributes to the international research community?

1. A technique that takes into account the individual properties of humans that contribute their domain knowledge to a machine learning process to perform the training of a classifier by making use of the idea of active learning.
2. A Python framework that incorporates this technique.
3. A user interface that can be used to gather input of human experts for image classification.
4. A mechanism to take into account the differences in the level of expertise among multiple human experts within an environment of human-in-the-loop machine learning.
5. A mechanism to determine the ideal number of human experts needed for efficient decision-making within a process of human-in-the-loop machine learning.

## 1. Introduction

Modern industrial systems, which are backed by AI in various fields of application, are currently progressing towards three primary objectives: sustainability, resilience, and, most importantly, *human-centricity*. Human-centricity underscores the need for taking humans into account

[*] Correspondence to: University of Natural Resources and Life Sciences Vienna, Peter Jordan Str. 82, A-1190 Wien, Austria.
*E-mail address:* andreas.holzinger@human-centered.ai (A. Holzinger).

https://doi.org/10.1016/j.jii.2025.100827
Received 7 September 2024; Received in revised form 3 February 2025; Accepted 8 March 2025
Available online 19 March 2025

in two distinct manners. When designing complex systems, it is important to consider multiple factors from a cross-disciplinary perspective [1,2]. This includes incorporating user-centered techniques [3], multidisciplinary system thinking, and most importantly, collaborative activities. This notion directly relates to human-centered AI, which aims to promote dependable, secure, and trustworthy approaches that uphold human values, rights, justice, and dignity [4–6]. The idea of steering AI towards a behavior that is in line with human values and intentions is referred to as *AI alignment*. The dangers of misalignment in AI include, for instance, untruthful answers of AI applications and the loss of control over AI systems [7,8].

AI received an almost unexpected hype in the last few years. Applications and tools that resort to AI-based systems raised great fascination not only in communities of people that are very familiar with topics of computer science, but also among manifold industrial environments. As a result, the pace at which new tools and models in the field of AI are developed and released is nowadays higher than ever, not only in consumer-oriented fields and industrial applications, but also in scientific environments [9]. New models and technologies outperform older ones in almost all possible benchmarks. These new technologies often reach great popularity among the public. One of the most popular applications that utilize methods of AI, ChatGPT, became the fastest-growing application for consumers, counting about 100 million monthly active users [10]. ChatGPT is a web-based service that aims at providing human-like conversation experiences to its users. The ability to provide these experiences is powered by AI, more precisely, by a large language model. Although this application was able to build up a fairly large amount of active users, the service is not without errors. The application is vulnerable to so-called hallucinations. Thus, ChatGPT, in certain cases, produces answers that seem to fit perfectly into the conversation but are actually untrue. This problem makes it impossible to use this service for critical applications. Other applications like Midjourney, that allow users to create realistic images from textual descriptions, and Google Bard, which also aims at providing conversation experiences that feel human-like, became also very popular among the public [9], but are also suffering from hallucinations [11].

For all these reasons, it is meanwhile recognized how important it is not only to consider humans, but to integrate humans in the design of complex digital systems and to incorporate human knowledge in the form of a human-in-the-loop [12,13]. Digital tools and simulation techniques can be used to model reality during the conception and design phase and to predict system behavior and human–system interaction, optimizing the design according to multidisciplinary requirements. A good example are digital twins. Although the terms digitalization, digital transformation and “digital twin” [14] imply automatism, the interaction of an industrial system with humans remains necessary, for which suitable user interfaces must be created that differ according to the various roles and user groups. A number of methods are used to develop such a system: among others, model-based system technology, system dynamics and user experience design. An industrial system harnesses the extended cognitive capabilities through cloud-managed artificial intelligence and semantic technologies such as knowledge graphs.

The rise of popular applications that utilize methods of AI would not be possible without high-quality training data. Obtaining training data that represents the ground truth in a way so that it covers all aspects of possible inputs and that it is free of artifacts is very often difficult because of the demanding process that is required to create a dataset of high quality that can be used for training in a Machine Learning (ML) scenario. This problem is referred to as the *data bottleneck* [15]. The samples in this dataset are very often gathered and preprocessed by human experts that have the respective domain knowledge to provide annotations for samples to be used in ML processes. For many applications, the role of the human within such processes remains static within the data gathering and preprocessing stages. However, the role of a human annotator within ML processes can, as it is the case with the role of humans within a Digital Twin, highly influence the costs and outcome of the whole operation [14]. Furthermore, taking into account the role of humans within processes of digital systems should be an integral part of Industry 5.0. This could be beneficial for the more efficient and more diverse utilization of human capital as it would reflect contemporary economic thinking [16].

The need for training data of high quality is supported by observations that show that the quality of the output of ML applications, especially large language models, significantly decreases with lower quality of training data [17]. Especially responsible for the output of low quality, for instance, hallucinations, of large language models are samples that are used for training from unreliable sources as well as biased training data [18].

In addition to that, new trends are emerging in various industries. These trends highlight that human-machine collaboration is becoming more and more important and promote the use of the workforce in a cooperative environment in which both machines and humans play a significant role, as numerous important tasks in an industrial environment still have to be managed manually [19]. Also, the tendency towards Industry 4.0 causes a shift in industrial engineering methods, away from traditional techniques and towards cyber–physical systems, which require adaptations made to the management of human resources and knowledge [20].

## 2. Background

The disagreement problem in machine learning, particularly in high-stakes domains such as healthcare and autonomous systems, poses significant challenges to both model reliability and user trust. This paper addresses this issue by examining it within the broader context of federated machine learning, human-in-the-loop systems, explainable AI in image classification, counterfactual reasoning, and the Rashomon effect. These areas provide essential background for understanding how model disagreements arise and persist, and how they relate to the quality of data used in training and evaluation.

Federated learning offers a decentralized approach where multiple models are trained across different data sources, often leading to inconsistencies in model predictions due to heterogeneous data quality and distribution. Human-in-the-loop frameworks aim to incorporate human expertise into the decision-making process, but disagreements between human inputs and model predictions are a frequent source of concern. In image classification, explainability becomes crucial in identifying the reasons behind model errors and disagreements, especially when conflicting interpretations are derived from the same input data. Counterfactual reasoning further enhances this by offering insights into how alternative scenarios or inputs would affect model outcomes. The Rashomon effect, which highlights the possibility of multiple valid interpretations for the same data, underscores the importance of addressing these disagreements.

The following background is necessary for understanding the relationship between model disagreements and data quality, as these fields contribute distinct perspectives on how poor or inconsistent data can exacerbate disagreements, ultimately affecting the robustness, transparency, and fairness of machine learning systems.

### 2.1. Federated machine learning

Processing data at a central instance has some critical disadvantages in terms of data privacy and load balancing, giving rise to the idea of a federated ML approach [21]. In contrast to the traditional approach where each participant that provides data to the ML process would have to submit this data directly to the central instance of the system, the idea of a federated ML process suggests that each participant neither exposes its data to the other participants of the network, nor to any central instance. Instead, participants work collaboratively on a global ML model. This technique usually involves the training of a separate model at each participant, and the combination of all local models by exchanging model updates to yield a globally learned ML model that maintains the privacy of the local data repositories [22].

In some federated ML scenarios, a central instance is used to orchestrate the process and to combine the results of the local training processes to a global model [23]. This usually works as follows: the central instance (server) broadcasts an initial model to all clients that participate in the federated ML process. Each client then performs a local training process using the model received from the server and updating it based on the local training dataset. The resulting changes to the model are then transferred back to the server. The server then aggregates all changes received by the clients to produce a new global model, which is then delivered downstream back to the clients for a new local update round [24].

Especially in the domain of medicine and health, AI/ML faces obstacles related to data privacy and the technical complexities of handling distributed data across multiple institutions. Consequently, federated learning resolves these concerns by facilitating the creation of ML models without the need for disclosing sensitive data. Nevertheless, current tools for federated learning and frameworks frequently lack adaptability, necessitate programming expertise, and are not easy to use for domain experts. Matschinske et al. (2023) [25] present a platform to democratize federated learning by offering a comprehensive solution for creating and implementing federated algorithms in the field of biomedicine and beyond. This platform streamlines the development and execution process for both developers and end users, whilst it guarantees data protection and compliance with prevailing legislation, including the European General Data Protection Regulation (GDPR).

Thus, to apply ML techniques to this data, it is very important to not expose data directly to other participants in a global network or to a central instance. Consequently, healthcare applications can make use of the idea of federated ML since it allows them to remain in control over their data. In this context, Pfeifer et al. (2023) [26] describe an approach to deploy federated ensemble-based Graph Neural Networks for healthcare applications, and Hausleitner et al. (2024) [27] introduce a framework that integrates federated learning with Graph Neural Networks to classify diseases, incorporating Human-in-the-Loop methodologies, where they use collaborative voting mechanisms on subgraphs within a Protein–Protein Interaction network, situated in a federated ensemble-based deep learning context.

In addition to that, federation plays a critical role in many industrial environments. A rising interest in distributed systems can not only be observed when it comes to traditional means of data processing but also when it comes to more advanced techniques such as Digital Twins [28].

### 2.2. The human in the loop concept

Human-in-the-loop refers to a model in which humans aid computers in decision making processes. In the area of ML, this refers to the interaction between a human and a ML process. In previous years, different techniques have been developed on how a human can interact with a ML process and how inputs of humans are used in such processes.

Mosqueira-Rey et al. (2023) [12] distinguishes three different approaches about how a human-in-the-loop can be integrated into a ML-system. The first approach is called Active Learning. This technique suggests that the ML-system completely remains in control over the learning process. The only task of the human-in-the-loop is to label and annotate data. They describe the human as an oracle for the ML-system which the system can consult to request the labeling and annotation of unlabeled data. Interactive ML describes an approach where humans interact more closely with the ML-process compared to Active Learning. This approach does not define one specific task for the human that is interacting with the ML-system. Instead, the idea of interactive ML proposes that at any time, human and computer should do whatever each of them does best. Ergo, the position of the human within the whole process is not static as it is in Active Learning. For instance, the human could go at the end of the ML flow to validate the correctness of results and perform adjustments if necessary. The third approach is called Machine Teaching. This idea describes the human as a teacher for the ML-system, thus, giving much more control over the whole process to the human-in-the-loop.

Furthermore, human-in-the-loop methods can be divided by the relationship of the human domain expert to the domain expert. *Before learning* describes techniques where the human expert resides before the actual training process such as *curriculum learning. During learning* describes methods that put the human expert directly in the training process. *After learning* puts the human expert at the end of the training process. This technique usually makes use of explainable AI. *Beyond learning* describes methods that go beyond the training process, such as *useful AI* [15].

An integration of a Human-in-the-Loop can bring several benefits to the information integration process. The Human-in-the-Loop (HITL) principle involves integrating human expertise directly into the decision-making processes of AI systems [29]. When applied in the context of explainable artificial intelligence (xAI), this principle allows experts to interact with AI models through a user interface (UI) [30]. This interaction is not merely passive but actively involves the expert in refining and guiding the AI's outputs. Consequently, experts can leverage this interaction to modify datasets in a way that produces counterfactual explanations [31]. Counterfactual explanations are hypothetical scenarios that help to understand how changes in the input data would lead to different outcomes from the AI model. By adjusting the data and observing how these changes impact the AI's predictions, the experts can gain insights into the decision-making process of the AI. This process also helps in validating the AI model's reliability and fairness, ensuring that the system's outputs align with human values and expectations. Moreover, this approach empowers experts to identify and mitigate potential biases in the AI model by experimenting with various data modifications. By providing counterfactual scenarios, experts can uncover hidden patterns or dependencies in the model, facilitating a deeper understanding of its behavior. This iterative process of human intervention, supported by explainable AI tools, ultimately enhances the transparency, accountability, and trustworthiness of AI systems.

The usefulness of such an integration of human expert knowledge in ML processes can be illustrated by applying it to real-world problems such as the classification of patients suffering from pancreatic cancer in terms of whether a chemotherapy treatment should be conducted [32].

Approaches like the integration of a human-in-the-loop are especially suitable for applications in the medical domain. This is because the availability of datasets that can be used for training in ML applications is reduced in the medical domain. Thus, traditional methods of training a classifier are not applicable due to insufficient training samples [33].

### 2.3. Explainable AI

Explainable AI (XAI) refers to methods and techniques in artificial intelligence that try to make the decision-making processes of AI systems re-traceable, hence transparent and consequently comprehensible to humans. Unlike traditional “black-box” models [34], where the internal workings are opaque and difficult to interpret, XAI aims to provide clear explanations of how an AI system arrives at its conclusions, predictions, or decisions [35]. The motivation for using such XAI methods is manifold:

1. Trust and Adoption: As AI systems are increasingly integrated into critical sectors such as medicine and healthcare, finance, and law, stakeholders need to trust these systems. Therefore the aim of XAI it to provide insights into the AI's decision-making process, enabling users to verify and validate the outcomes, thereby fostering trust and wider adoption.

2. Accountability and Compliance: Regulatory frameworks, such as the GDPR in Europe, require that decisions made by automated systems can be explained, especially when they significantly impact individuals. XAI helps organizations comply with these regulations by providing the necessary transparency.
3. Ethical Considerations: AI systems can sometimes produce biased or unfair outcomes. XAI enables the identification and mitigation of such biases, ensuring that AI systems operate ethically and do not perpetuate or exacerbate existing inequalities.
4. Improved Decision-Making: For experts using AI as a tool in their decision-making processes, understanding the rationale behind an AI's recommendation is crucial. XAI allows these users to critically assess AI outputs and make informed decisions.
5. Safety and Reliability: In applications where AI systems operate autonomously, such as self-driving cars or autonomous drones, explainability is essential for diagnosing errors, improving system design, and ensuring overall safety and reliability.

Various techniques implement the ideas of XAI. Among them, the idea of counterfactuals is one of the most popular methods of XAI. Counterfactuals, in terms of XAI, describe the altered input vector which results from applying changes to the original input vector in order to change the output of the model. Moreira et al. conducted investigations on the generation of counterfactual explanation by performing an evaluation on different types of models [36].

### 2.4. Explainable AI in image classification and counterfactuals

As stated above, one goal of XAI is to provide explanations on how different features of the input data are contributing to the output of an AI system. In the area of image analysis, the question about which attributes contribute to the output basically translates to the question asking for the regions of an image that contribute to a certain output. Saliency maps try to give answers to exactly this question by giving a visualization of the areas of an image that are of special importance for the computed output [36,37].

To obtain a saliency map that shows the regions of an image that are important for the output of a certain AI tool, it is crucial to think about how such regions can be computed from the image. These regions can be obtained by changing parts of the image and observing which changes in the image lead to changes in the output of the classifier. The regions that, when altered, lead to a change in the output, are therefore of special interest. The image that results from changing these regions in the original image is then called the *counterfactual* image. The resulting class from applying the image classification technique to the *counterfactual* image is referred to as the *counterfactual* class [38].

As stated above, one goal is trust and adoption. A currently very topical aspect is that trust can be achieved through counterfactual explanations. In this way, people can familiarize themselves with unknown processes by understanding the hypothetical input conditions under which the outcome changes [39]. This allows users to understand the logic behind an AI system's decisions by seeing how alternative scenarios could lead to different outcomes. This not only promotes understanding, but also strengthens the ability to anticipate the AI's behavior and identify potential sources of error. In practice, this contributes significantly to acceptance of and trust in the technology, as the decision-making processes can be made comprehensible and transparent.

There are different methods of altering the original image to trick an image classifier into outputting the *counterfactual* class. These methods include removing or blurring parts of the original image but also adding features that are associated with the desired *counterfactual* to the image [40]. However, the approach of adding concepts to an image to trick the classifier into assigning a *counterfactual* class to the image is criticized by Vermeire et al. as this approach would lead to counter-intuitive explanations that are not useful when trying to understand how and why a certain assignment to a class was made by an image classifier. Thus, Vermeire et al. (2022) [38] recommend to avoid the addition of evidence to images in order to generate useful *counterfactual* images.

Furthermore, Vemeire et al. (2022) propose and summarize different techniques to perform a segmentation of the original image into regions as well as techniques to remove evidence within a certain region from the image, whereas the most suitable choice for segmentation turned out to be *quick-shift* [41] and the best choice for evidence removal turned out to be blurring certain regions by using Gaussian Blur [38].

### 2.5. The rashomon effect and the disagreement problem in the context of XAI

The disagreement problem and the Rashomon effect are distinct concepts in the context of explainable AI, but they both relate to the multiplicity of explanations or models. The disagreement problem arises when multiple models produce different explanations or decisions based on the same data or inputs. This can occur in AI systems where different models, possibly trained on the same dataset, reach different conclusions due to variations in their architectures, training processes, or inherent randomness. The disagreement problem is particularly relevant in ensemble methods or when comparing models trained independently. It raises concerns about the consistency and reliability of AI systems, especially in critical applications where different models should ideally converge on similar decisions. The Rashomon effect refers to the phenomenon where multiple plausible models or explanations exist for the same dataset or scenario. In the context of XAI, the Rashomon effect highlights the fact that multiple models can fit the data well, yet provide different interpretations or explanations of the underlying relationships. This effect underscores the idea that there is often no single "correct" model, but rather a set of models that are equally valid according to the data, leading to challenges in model selection and interpretation.

The Rashomon effect specifically refers to the phenomenon where different people have varying and contradictory interpretations of the same event due to differences in perspective, memory, and personal biases. This effect is named after Akira Kurosawa's 1950 film "Rashomon", which depicts multiple characters providing different accounts of the same incident. The key characteristics of this effect are:

- Subjectivity: differences arise from subjective perceptions, memories, and biases.
- Narrative Variability: Multiple, often conflicting, narratives about a single event.
- Contextual Focus: Often discussed in the context of eyewitness accounts, storytelling, and historical interpretation.

In the context of AI, the Rashomon effect has been described as a phenomenon in which different explanations of ML are obtained when different models are used to describe the same data. This serious and growing problem is an emerging topic in the explainable AI community [42].

The disagreement problem, in a more general sense, refers to the challenge of resolving differences of opinion or belief between individuals or groups. This problem is widely discussed in philosophy, particularly in epistemology, ethics, and social theory. Unlike the Rashomon effect, the disagreement problem focuses on examining the variations across various counterfactual explanation algorithms applied to the same event and classifier. Nevertheless, the crux of the issue lies in the fact that both the Rashomon effect and the dispute problem present identical ethical concerns and moral risks: the question of who has the authority to determine which explanation will be adopted [43].

Key characteristics of the disagreement problem are, for instance, epistemic differences which arise from differences in knowledge, evidence, or reasoning, as well as the resolution Focus that deals with methods and principles for resolving or understanding disagreements, such as the role of evidence, rationality, and dialogue. Moreover, it is important to note that the disagreement problem is relevant in many areas including scientific debates, moral disagreement [44], legal disagreement [45] and medical decision disagreement [46].

Thus, the Rashomon effect originates from subjective perception and memory, leading to differing accounts of the same event. In contrast, the disagreement problem originates from differences in knowledge, evidence, or reasoning, leading to differing opinions or beliefs. While the Rashomon effect is typically discussed in contexts involving personal narratives and interpretations, the disagreement problem is broader, encompassing various forms of intellectual, ethical, and practical disagreements. Therefore, the Rashomon effect highlights the unreliability of subjective accounts and the complexity of human perception, and the disagreement problem focuses on the challenges of achieving consensus or understanding in the face of differing views.

As different post-hoc explanation methods are increasingly used to explain complex models in high-risk situations, it becomes increasingly important to develop a deeper understanding of if and when the explanations issued by these methods diverge and how such disagreements are resolved in practice. Krishna et al. (2022) [47] formalize the notion of disagreement between explanations and analyze how often such disagreements occur in practice and how practitioners resolve these disagreements. ML practitioners often apply ad hoc heuristics.

Recent XAI research has shown that current model-independent counterfactual algorithms for explainable AI are not based on a causal theoretical formalism and therefore cannot promote causality to a human decision maker. Furthermore, new results suggest that the explanations derived from common algorithms in the literature provide spurious correlations rather than cause-and-effect relationships, resulting in suboptimal, flawed, or even biased explanations [48].

The XAI disagreement problem concerns the fact that different explainability methods provide different local/global insights into model behavior. Since there is no certain truth in explainability, practitioners ask themselves: ‘Which explanation should I believe?’ In a very recent paper, Laberge et al. (2024) [49] approached this problem from the perspective of functional decomposition. Many XAI techniques do not agree because they treat feature interactions differently.

## 3. Methods

### 3.1. Performance metrics

To efficiently optimize the amount of needed consultations and because each human annotator performs differently, a detailed evaluation of the annotator’s performance is important during the process. The calculated performance metrics are used to ensure that the resulting annotation, that is then used to train a machine classifier, is of high quality.

Usually, performance metrics of a classifier are calculated after the training process is finished. For binary classification, many commonly known performance metrics, such as the F-measure, build on the idea of the confusion matrix. This matrix displays the amount of true positives, true negatives, false positives and false negatives of a certain classifier. Ergo, to compute these four values, the ground truth has to be known [50].

#### 3.1.1. Binary classification

Binary classification is a special form of classification where one out of two classes is assigned to each sample. To be precise, binary classification tests are usually used to detect the presence of a certain condition. The two possible classes of the binary classification task then denote either the presence or the absence of such a condition. The performance of binary classifiers is measured using a reference test, the so called *gold standard*. This *gold standard* serves as ground truth to determine if predictions of the classifier are correct or not. A prediction can be categorized, in terms of the classifier, into four possible outcomes. It is either true positive, false positive, true negative or false negative. The performance of the binary classifier is then often given as sensitivity and specificity. Sensitivity is defined as $\frac{a}{a+c}$ and specificity is defined as $\frac{d}{c+d}$, where $a$ is the number of true positives, $b$ is the number of false positives, $c$ is the number of false negatives and $d$ is the number of true negatives [51].

However, the techniques to determine the performance of a binary classifier mentioned above require data from the *gold standard*. Ergo, to measure the performance of such a classifier, the ground truth has to be known. The precise estimation of specificity and sensitivity of a binary classifier can be difficult, as highlighted by Keddie et al. since various estimation techniques tend to deliver biased estimations. This can be dangerous when it comes to evaluating the usefulness of a certain binary classifier. A underestimation of the classifier’s performance could lead to the discarding of a per se acceptable classifier and a overestimation of the performance could cause relying on a bad classifier [52].

The data that serves as the *gold standard* often originates from manual classification of samples. This manual classification task is usually done by humans. However, human interaction with data is not always desirable, mainly because it is very expensive. To overcome this issue, Tripathi et al. proposed a sampling method to estimate the performances of multiple classifiers that makes use of overlaps between the prediction sets. Using this sampling method, Tripathi et al. were able to reduce the amount of needed annotations done by humans to form the *gold standard* to accurately estimate the performance of multiple binary classifiers [53].

#### 3.1.2. Multi-class classification

For scenarios that are beyond binary classification, such as classification processes with three or more possible classes, computing the performance metrics mentioned above is possible anyway by breaking down the classification problem into binary classification problems. Furthermore, certain performance metrics can be calculated directly. These metrics include the overall agreement rate, defined as $overall AR = \frac{\sum a_i}{T}$, where $a_i$ denotes the number of agreements for class $i$, ergo the number of correctly classified samples for class $i$, and $T$ denotes the overall amount of classifications, as well as the overall error rate, which is defined as $overall ER = 1 - overall AR$ [54].

However, the performance metrics mentioned above, per definition, also require the knowledge of the ground truth in order to assess the correctness of a classification. Since in many real-life scenarios, this ground truth is not known in advance, a different metric to measure the performance of a human annotator has to be found. To be precise, the definitions of the performance metrics have to be adjusted to make use of a value other than the ground truth or a *gold standard*. Ergo, an alternative to the ground truth has to be found.

## 4. Notations

To efficiently describe the ideas and findings in this work, it is crucial to define some important notations. These notations are used throughout the remaining parts of this work. A binary classifier is defined as a function $C$ that assigns a label $\in \{0,1\}$ to samples from a universe $\Omega$. The function $C$ for a sample $x$ is therefore defined as $C : x \mapsto \{0,1\}$ [53,55].

Thus, a classifier that has more than two possible outcomes, ergo is beyond binary classification, is then denoted as $C : x \mapsto \{0 \dots n\}$.

For certain aspects of this work, this definition can be extended to not only include the output as an assignment of a positive or negative label, but also include an output value that we can interpret as a score that indicates the probability that the sample belongs to either the positive or the negative class. The resulting definition, again for a sample $x$, can be written as $C' : x \mapsto y$, where $y$ can be an arbitrary numerical value. For the outcome of $y$, we can define a threshold $t$ to

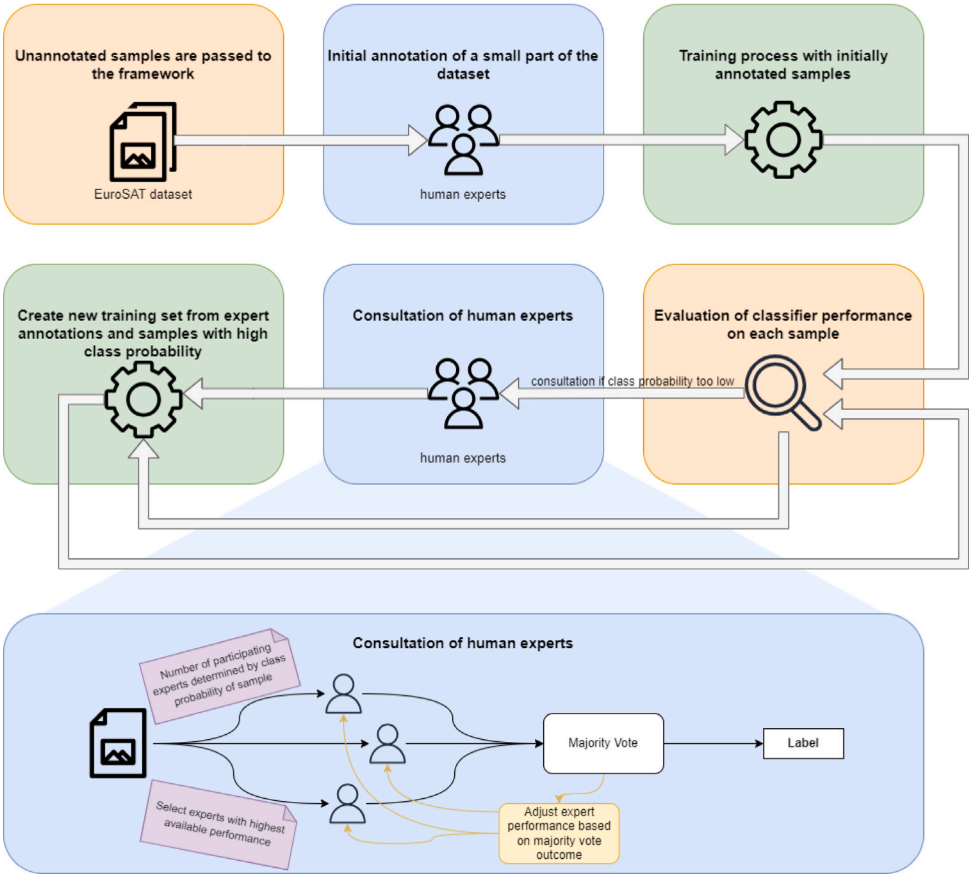

Unannotated samples are passed to the framework
EuroSAT dataset
Initial annotation of a small part of the dataset
human experts
Training process with initially annotated samples
Create new training set from expert annotations and samples with high class probability
Consultation of human experts
consultation if class probability too low
human experts
Evaluation of classifier performance on each sample
Consultation of human experts
Number of participating experts determined by class probability of sample
Select experts with highest available performance
Majority Vote
Label
Adjust expert performance based on majority vote outcome

**Fig. 1.** Our proposed approach to train a classifier in a human-in-the-loop environment.

map samples to either the positive or negative class using the values of $y$, whereas a value $y < t$ causes the sample to be assigned a positive label. Otherwise, the negative label is assigned [55].

It is assumed that always $y \in [0,1]$. For the sake of this work, it is assumed that $t = 0.5$. Thus, we assume no imbalance between the probabilities of the individual possible classes, to be precise the positive and negative class. Ergo, a sample is assigned to the positive class if $y < 0.5$ and to the negative class if $y \geq 0.5$.

For non-binary classifiers, the output can, in certain cases, also be interpreted as a score that indicates the probability of a sample belonging to a certain class. For all possible outcomes, the sample is considered a member of the class showing the highest class probability.

Furthermore, we define some additional properties that are important for the behavior of our system. We denote $M$ as the set that contains data to train the classifier. Ergo, the amount of samples that are available for the training process is limited. The $i$th sample of $M$ is denoted as $M_i$. The amount of samples in $M$ is denoted as $m$. We denote the set that contains the ground truth and serves as a *gold standard* to train a reference classifier as $GT$, whereas the $i$th item of $GT$ is called $GT_i$. As $GT$ contains the ground truth, to be precise, the correct label for all samples in $M$, the correct label for $M_i$ is $GT_i$. We also define the set $N$ that holds the human experts that provide knowledge to the system. These humans are modeled, in the case of a binary classification problem, as binary classifiers with each of them having its specificity and sensitivity. We denote the $i$th human expert in $N$ as $N_i$ and the amount of human experts in $N$ as $n$. Each human expert that is modeled as binary classifier here, can be seen as, in mathematical terms, a function, whereas $N_i : x \mapsto \{0,1\}, x \in M$. Again, 0 denotes the positive class and 1 denotes the negative class. Furthermore, we denote the true, but unknown, specificity and sensitivity of the $i$th human expert as $Sp_i$ and $Se_i$. Since the true sensitivity is not known to the system, we denote the estimated specificity and sensitivity of the $i$th human expert as $Sp'_i$ and $Se'_i$. Thus, the absolute estimation error for the $i$th human expert equals $|Sp_i - Sp'_i|$ and $|Se_i - Se'_i|$. For a problem that goes beyond binary classification, we measure the overall agreement rate $overallAR$. This metric defines the performance of the classifier or human expert. The true but unknown performance of the $i$th human expert is therefore denoted as $p_i$ and the observed performance is denoted as $p'_i$.

## 5. Procedure

With this work, we aim at developing methods to optimize the utilization of human interaction needed in order to efficiently train a (binary) classifier by making use of the beneficial properties of a human in the loop. To reach this goal, we stick to the idea of active learning where the system consults humans to incorporate the expert knowledge provided by this human expert into the training process of the classifier. Our proposed approach is visualized in Fig. 1. We assume that at the beginning of the training process, no annotated data is available and the only way for the system to obtain annotated data is through consultation of human experts. These human experts provide input to the system in the form of annotated samples. Thus, each human expert can be modeled as a (binary) classifier, whereas each expert has its own specificity and sensitivity in case of binary classification. For other cases of classification, the performance is measured by making use of the $overallAR$. It is assumed that the true performance metrics are not known to the system. Thus, the performance of each human expert has to be evaluated during the training process.

The goal of the system is to output a classifier that was trained using a minimal amount of consultations of human experts. To reach this goal, the quality of the trained classifier would have to be evaluated against another classifier. This other classifier should serve as a *gold standard* for the purpose of evaluating the output of the system. The classifier that serves as the *gold standard* here is trained using a dataset for which all samples are already annotated. To determine the differences in quality between the classifier trained by our system and the *gold standard* classifier, both classifiers are tested against a special, already annotated, dataset.

### *5.1. Preliminary observations for binary classification*

To efficiently make use of the aspect of including a human-in-the-loop in the training process of a binary classifier, it is important to accurately measure the performance of individual humans taking part in the process. This is done to determine the value of an input made by a specific human. This value is essential to form a technique to reduce the number of human consultations during the training process of a binary classifier.

To accurately represent the performance of a human-in-the-loop, it is necessary to estimate its performance metrics, such as the sensitivity and specificity in the case of binary classification. However, as mentioned above, to calculate such metrics, a *gold standard*, that is used a reference to compare the underlying (human) classifier, is needed. Ergo, already annotated samples are required in order to compute the desired performance metrics. An intuitive way to substitute the non-existent *gold standard*, is to present the same unannotated sample to multiple human experts. Then, perform a majority vote on the received responses and treat the result, ergo, the consensus among all the consulted human experts, as the temporary *gold standard*.

It is common sense that an estimation, as mentioned above, would require a bigger amount of datapoints from the behavior of the individual consulted human experts to reach a reasonable level of accuracy of the estimated performance metrics. To get an idea of the required number of consultations of an individual human expert to estimate its performance, ergo its specificity and sensitivity, with satisfying accuracy, a preliminary experiment was conducted. The goal of this preliminary experiment was to get an idea of how accurate an estimation of sensitivity and specificity of a binary classifier can be without being dependent on the existence of a *gold standard*.

For this experiment, random data was generated. A dataset of 10000 samples was generated that served as the ground truth in this experiment whereas data was generated so that 75% of the samples are members of the negative class and 25% of the samples are members of the positive class. Furthermore, $k$ primitive binary classifiers were generated, whereas we tested the following configurations for $k$: $k \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Each synthetic classifier was generated with a random specificity and sensitivity. The randomly generated sensitivity was in the range $[0.7, 0.9]$ for every generated classifier and the randomly generated specificity was in the range $[0.8, 0.98]$. In general, the results of this experiment show that with an increasing amount of participating classifiers per decision, the error of the estimation of the sensitivity and the specificity of an individual classifier decreases. In other words, the more classifiers are participating in the decision-making process, the higher gets the accuracy of the estimation of sensitivity and specificity of those classifiers.

To illustrate some of the result of this observation, Fig. 2 shows the estimated sensitivity and specificity for the first 1000 classifications for different numbers of involved classifiers. For the sake of readability, not all numbers of involved classifiers are shown in the figure. To be precise, the data for 2, 3, 6 and 10 classifiers are shown in combined charts in Fig. 2. All other individual charts for this observation can be found in the appendix.

These observations give an idea about the reliability of estimations of the performances of binary classifiers. They do not only indicate that with a rising number of participating classifiers the average error of an estimation of the sensitivity and specificity decreases, but also that with a rising number of samples seen by the classifiers, the error of the estimations also decreases. Furthermore, the error of the estimations is very high when the amount of samples seen by the classifiers is very low or the amount of participating classifiers is very low.

### *5.2. Principles*

Based on the preliminary observations and previous work, we want to propose a technique to optimize the utilization of human annotators in a classification process to efficiently train a classifier. To define this technique, a few important principles and hyperparameters have to be defined.

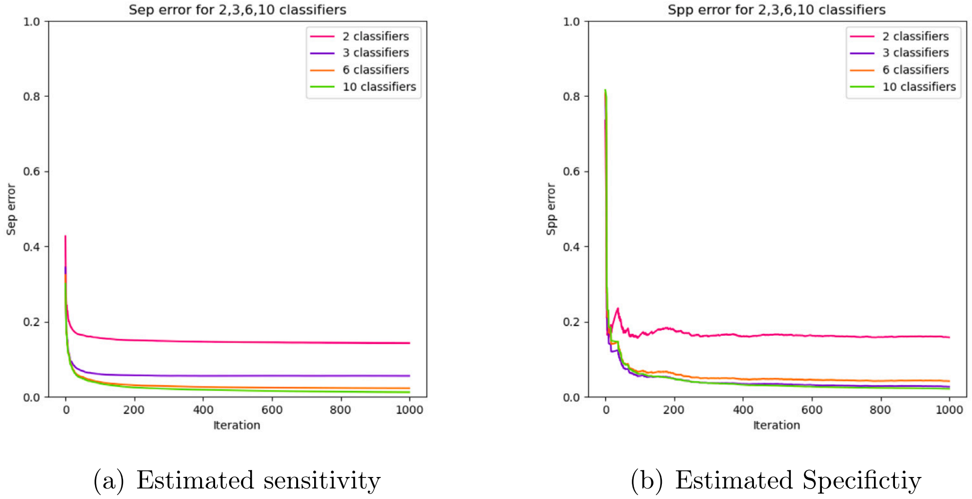
In the beginning of the process, the only input is a dataset of samples that have to be classified, where each sample is a member of exactly one class. Since at the beginning of the process, no data is classified, a specific amount of samples have to be classified by human annotators. Since also no data about the performance, ergo, the sensitivity and specificity or the overall agreement rate, of human annotators is known at the beginning, human annotators can be chosen randomly from the pool of available annotators. Furthermore, as mentioned above, the error of the estimation of performance metrics is very high at the beginning of the process. Thus, the estimated values for sensitivity and specificity for each human annotator are unreliable in the initial phase of the process.

## 6. Implementation

Based on the above findings, we propose a framework that can be used to optimize the utilization of humans in the loop within a specific ML process, more precisely, within a process of Active Learning. The goal of this framework is to reduce the number of required human interactions while still maintaining a high level of quality of the trained classifier.

The framework assumes that no labeled data is available at the beginning of the process. Thus, a certain portion of the dataset containing the samples that have to be classified has to be labeled at the beginning. This initial labeling is also handled by the framework. It is also assumed that multiple human annotators take part in the process and that the level of expertise is not the same for all annotators. Thus, it is assumed that each human annotator has its individual trustworthiness, ergo, its individual performance. This trustworthiness is evaluated and adjusted during the process by comparing annotations given by one annotator to those given by all other annotators that take part in the respective annotation round. The trustworthiness is represented in the whole process as the performance of the annotator. However, since the trustworthiness of each human annotator is calculated using the previously given annotations of the respective annotator compared the annotations given by all other annotators, a precise value for the trustworthiness cannot be computed at the beginning of the process. Thus, for a certain portion of all annotation tasks, the trustworthiness cannot be used to determine the quality of an annotation performed by a human annotator.

The framework is realized as Python application that orchestrates and performs certain tasks within an environment of Active Learning for the user. However, some aspects of the process are still in the responsibility of the user. An illustration showing the responsibilities of the user as well as those of the framework is shown in Fig. 3. At the beginning of the workflow, the framework performs an initial consultation of annotators to obtain an initial annotated dataset that can be used for the first round of training a classifier. Then, the resulting classifier is tested against the whole unannotated dataset to obtain the class probabilities for each sample. Depending on the class probabilities that are returned by the current classifier, samples are selected for the consultation of a human expert. Usually, a sample is presented to more than one human expert to obtain multiple annotations for the same sample. The resulting annotation is then obtained by performing a majority vote. The number of human experts that are consulted for one specific sample is determined by the class probability of the respective sample. Since a greater class probability means a greater confidence of the classification made by the current classifier, less human experts



(a) Estimated sensitivity

(b) Estimated Specifictiy

**Fig. 2.** Chart showing the error of the estimated sensitivity (a) and specificity (b) for the first 1000 classifications using the approach described above with 2, 3, 6 and 10 classifiers per decision.

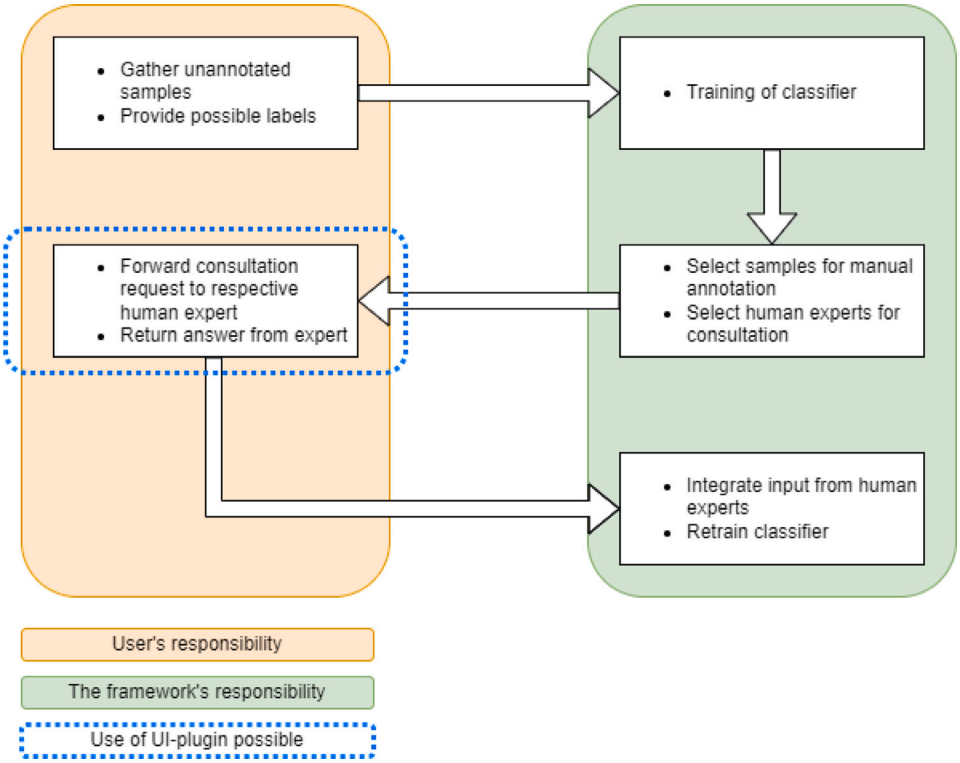

Gather unannotated samples
Provide possible labels
Training of classifier
Forward consultation request to respective human expert
Return answer from expert
Select samples for manual annotation
Select human experts for consultation
Integrate input from human experts
Retrain classifier
User's responsibility
The framework's responsibility
Use of UI-plugin possible

**Fig. 3.** Responsibilities within the proposed framework.

are consulted if a classification attempt shows a greater class probability. Samples that show a very high class probability are considered annotated without consultation of a human expert. The exact thresholds can be set by the user when calling the framework. The performance of each annotator is constantly evaluated by measuring the overall agreement rate. This is done by comparing the number of consultations done and the number of correctly annotated samples by the means of the result of each majority voting. For each consultation round, the human experts with the highest measured performance are selected for consultation. However, since the performance measurement can be very imprecise due to the low amount of available datapoints, the performance measurements are not considered at the beginning of the workflow. Instead, human annotators are chosen randomly. In addition to that, the available capacity, ergo, the maximum number of consultations that one specific human expert can perform, is considered when selecting human experts for consultation. After the consultations, the returned annotated samples are merged with all other samples that are already annotated and the classifier is now trained with the expanded dataset. This process repeats until one of various possible conditions is met. The process is stopped, and the final classifier is returned when either, a certain performance on a test dataset is met, a certain number of annotation rounds is met or the is not enough capacity for human consultations among all human experts. In addition to that, the framework features a user interface to simplify gathering the input of human experts. This user interface allows human experts the examination of images as well as sending their input to the framework. The user interface is a web-based interface that was developed using the *Flask* framework for Python. Fig. 6 shows the user interface displaying

```
# annotators: list of instances of the Annotator
# class
# a: the fraction of the dataset that is used for
# initial classification, e.g. 0.05
# call_annotator: a function that gets called when
# an annotation of a specific annotator is requested
# prob_high: upper class probability threshold
# for human annotation
# prob_low: lower class probability threshold
# for human annotation
# training_data: the unannotated samples
# possible_labels: potential labels to be assigned
# to the dataset
# min_annotators: the minimum number of
# human experts taking part in one annotation round
# max_annotators: the maximum number of
# human experts taking part in one annotation round
# test_data_X: the test set
# test_data_y: the labels of the test set
# additional_data: additional data passed to
# call_annotator, must be of same size as
# training_data
HITLAnnotator(annotators,
              a,
              call_annotator,
              prob_high,
              prob_low,
              classifier,
              training_data,
              possible_labels,
              min_annotators,
              max_annotators,
              test_data_X,
              test_data_y,
              additional_data)
```

**Fig. 4.** The signature of the constructor of the HITLAnnotator class.

an image at a resolution of 64 by 64 pixels from the EuroSAT dataset with four possible classes that the image could be assigned to.

The framework can be imported via the *HITLAnnotator* class. This class is the main entry point for all interactions with the framework. The signature of the constructor is shown in Fig. 4. The argument *call_function* is of special importance here because This argument is a function that gets called when a consultation of a human expert is requested. Furthermore, the framework allows the import of the Annotator class. This class represents one single human expert. The signature of the constructor is shown in Fig. 5. In "primitive" mode, a class is returned randomly based on the parameter's *accuracy_high* and *accuracy_low* and the resulting accuracy. Ergo, the computed accuracy defines if the correct class is returned or an incorrect class. Thus, the mode "primitive" can be used for simulation purposes.

The source code of the framework is available at https://github.com/human-centered-ai-lab/app-HITL-annotator.

## 7. Evaluation and discussion of results

We present a framework to optimize the utilization of human expert knowledge within a process of ML. This is done by making use of Active Learning and by measuring the performance, ergo, the trustworthiness, of human experts during the whole procedure. To measure the success of our framework, we conducted a test of the framework on the EuroSAT dataset. This dataset consists of 27000 labeled aerial images. The images belong to one of 13 classes [56]. To find useful parameters for a comparison, we tested various combinations of inputs. We compared the number of requested human consultations to the size

```
# id: id of the annotator, must be unique
# mode: 'function' or 'primitive', function:
# call_function will be called when annotation is
# requested, primitive: a label will be returned
# based on accuracy_high and accuracy_low
# accuracy_low: lower boundary for
# randomly generated accuracy of primitive
# classifier
# accuracy_high: upper boundary for
# randomly generated accuracy of primitive
# classifier
Annotator(id,
          mode,
          limit,
          accuracy_low,
          accuracy_high)
```

**Fig. 5.** The signature of the constructor of the Annotator class.

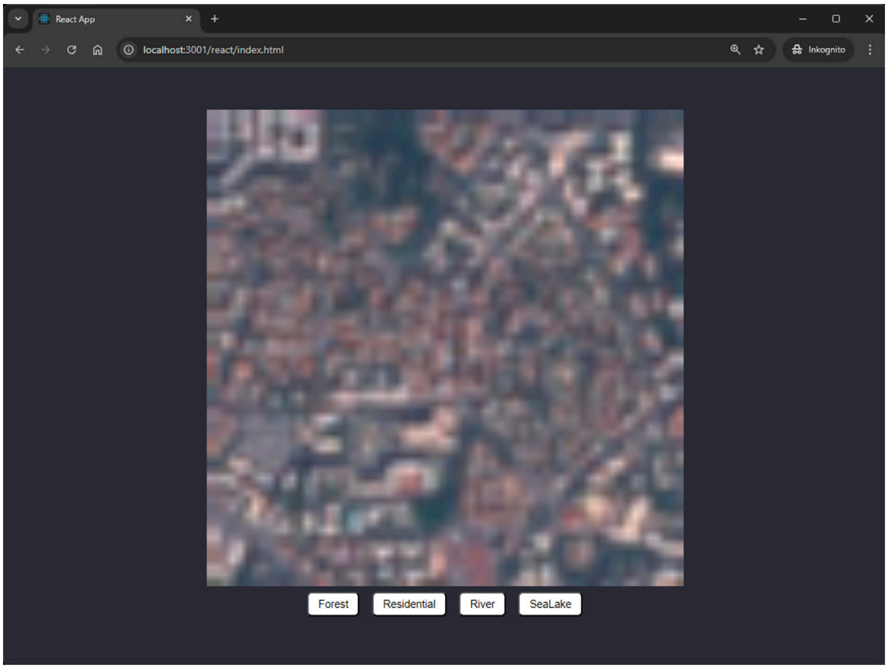


**Fig. 6.** The user interface.

of the test dataset, assuming that one human interaction was required to annotate one sample in the test dataset. For all of our test, the same image classifier from the *sklearn* Python package was used. To test the framework, we assumed an environment where 20 human experts are taking part in the process. These experts were modeled using the *Annotator* class within the Python framework. To perform a simulation, the *mode* parameter was set to "primitive". Each virtual human expert had a randomly generated accuracy. This accuracy lies in the range between 0.75 and 0.99. These values were chosen based on the fact that we assume human experts to have certain expert knowledge, resulting in a higher accuracy when annotating samples requiring exactly this expert knowledge. To give an idea of the effectiveness of the proposed framework, we measure the number of required human interactions in relation to the size of the dataset and compare the accuracy of the resulting classifier to the accuracy of a classifier that was trained using a completely annotated dataset. For this test, we used a subset of the EuroSAT dataset. This subset was generated by selecting those samples from the dataset that belong to the classes *River*, *SeaLake*, *Forest* and *Residential*. The dataset was split into a training set and a test set whereas 80% of the samples were randomly assigned to the training set and 20% of the samples were assigned to the test set. Testing showed that the overall performance is highly dependent on the chosen parameters. However, the results show that the number of required human interactions could be significantly decreased.

Furthermore, we applied the proposed framework to a different dataset, namely CIFAR-10 [57]. This dataset consists of 60000 images that belong to exactly one of ten possible classes. We applied the proposed framework to a subset of the CIFAR-10 dataset. For this subset, 10000 images were randomly selected. For evaluation of the framework using the CIFAR-10 dataset, the same configuration as for the evaluation using the EuroSAT dataset was used. In contrast to the configuration above, we only changed the portion of samples to be classified initially to 0.02 and the threshold for the consultation of a human to the range 0.5 to 0.55.

To put the measured performances in relation, we compared the results to a classifier that was trained the traditional way. For this benchmark scenario, we assumed that the samples that are used for the training process are annotated by exactly one human annotator per sample. The result of this annotation is then treated as the *ground truth* for respective sample. Ergo, the number of required human interactions equals the number of samples in the process. Using the framework on the EuroSAT dataset with the parameters set to allowing the participation of 1 to 3 human experts per decision and the fraction of the dataset to be annotated initially set to 0.1 as well as the upper limit of the class probability to trigger a consultation of a human expert to 0.6 and the lower limit to 0.5, resulted in a classifier that showed an accuracy of 82% whereas the classifier that was trained using the already annotated dataset reached an accuracy of 86%. However, the classifier that was trained using the proposed framework, required 42% less human interactions, assuming that to create the dataset in a traditional way, one interaction per sample is necessary. And using the framework on the Cifar-10 dataset resulted in a number of required human interactions that was 19% lower than for a training process that worked the traditional way, while delivering only slightly lower performance. While the proposed techniques may not drastically improve the performance of the resulting classifier, we proposed a technique to enable HITL machine learning in a scenario where different experts bring in their domain knowledge and the goal is to maximize the utilization of this highly specialized domain knowledge The proposed framework, in theory, allows the experts to be a different locations. We were able to show that the better utilization of human domain knowledge through our proposed framework results in a lower number of required human interaction within a machine learning process.

## 8. Future work

A framework was proposed to perform the training of a classifier within a process of active learning in a user-centric manner. The proposed framework takes into account the individual level of expertise of each human expert taking part in the process. This leads to in improved utilization of human expert knowledge when it comes to training a classifier using manually annotated data. The system is designed to allow a great range of customization. For instance, a custom function that is passed to the framework and gets called whenever the knowledge of a human expert is needed allows maximum customizability when designing the workflow for gathering manually annotated data. Furthermore, the proposed framework features a user interface to easily collect annotations from human experts that is designed as an add-in for the framework. However, as the main purpose of this implementation is to show its usefulness and not to run in a production environment, the system does not feature any authentication and encryption capabilities for the communication between the server and the clients.

Furthermore, the accuracy of the system could be improved by introducing methods of explainable AI. Explainable AI could be used to provide human experts with suggestions for the correct class as well as explanations on how these suggestions were generated. One method of generating such explanations are counterfactuals [37]. In the case of image classification, such explanations could be generated using counterfactual images. On the side of the human expert, these explanations could be visualized using saliency maps showing the most important regions of an image for a specific suggestion. Saliency maps could also be integrated directly into the already existing user interface.

While this work makes use of an image classification problem to showcase the usefulness and effectiveness of the proposed framework, the framework can, in theory, be applied to any classification problem. However, it is important to point out that the application of this framework on other types of classification problems still needs further testing and validation. The application of the proposed framework to different types of classification problems is intended to be uncomplicated as the handling of individual samples within the process falls under the responsibility of the user. To be precise, the code to handle an individual sample has to be brought in the by the user and gets called by the framework when necessary.

## 9. Conclusion

In times where applications that utilized the capabilities of ML are more popular than ever, the need for high quality training data for ML processes is also rising. With a great number of datasets that are used for ML being still generated using manual annotation, the need for a framework that allows to perform these kinds of annotations in a user-centric manner is of great importance. We propose such a framework that tries to optimize the utilization of human experts by making use of the fact that the level of expertise and domain knowledge varies among the human experts. This is done by constantly measuring the accuracy of each human expert and selecting experts for annotation tasks based on these measurements. We also show the usefulness of the proposed framework by evaluating it on the EuroSAT dataset that features 27000 aerial images of Europe. Furthermore, we give ideas for the extension of the existing frameworks such as the integration of state-of-the art techniques of explainable AI to give suggestions to human experts as well as explanations for those suggestions. We believe that such a framework can contribute to a better utilization of human experts that are willing to make their knowledge available to ML processes. This can especially be the case for environments where extensive domain knowledge is rare and expensive, such as medical applications as well as complicated industrial environments. For instance, the classification of medical images could benefit from such a framework because of the better utilization of the knowledge of domain experts.

## CRediT authorship contribution statement

**Matthias Huelser:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation. **Heimo Mueller:** Writing – review & editing, Funding acquisition, Data curation, Conceptualization. **Natalia Díaz-Rodríguez:** Writing – review & editing, Validation, Supervision. **Andreas Holzinger:** Writing – review & editing, Writing – original draft, Investigation, Funding acquisition, Conceptualization.

## Declaration of competing interest

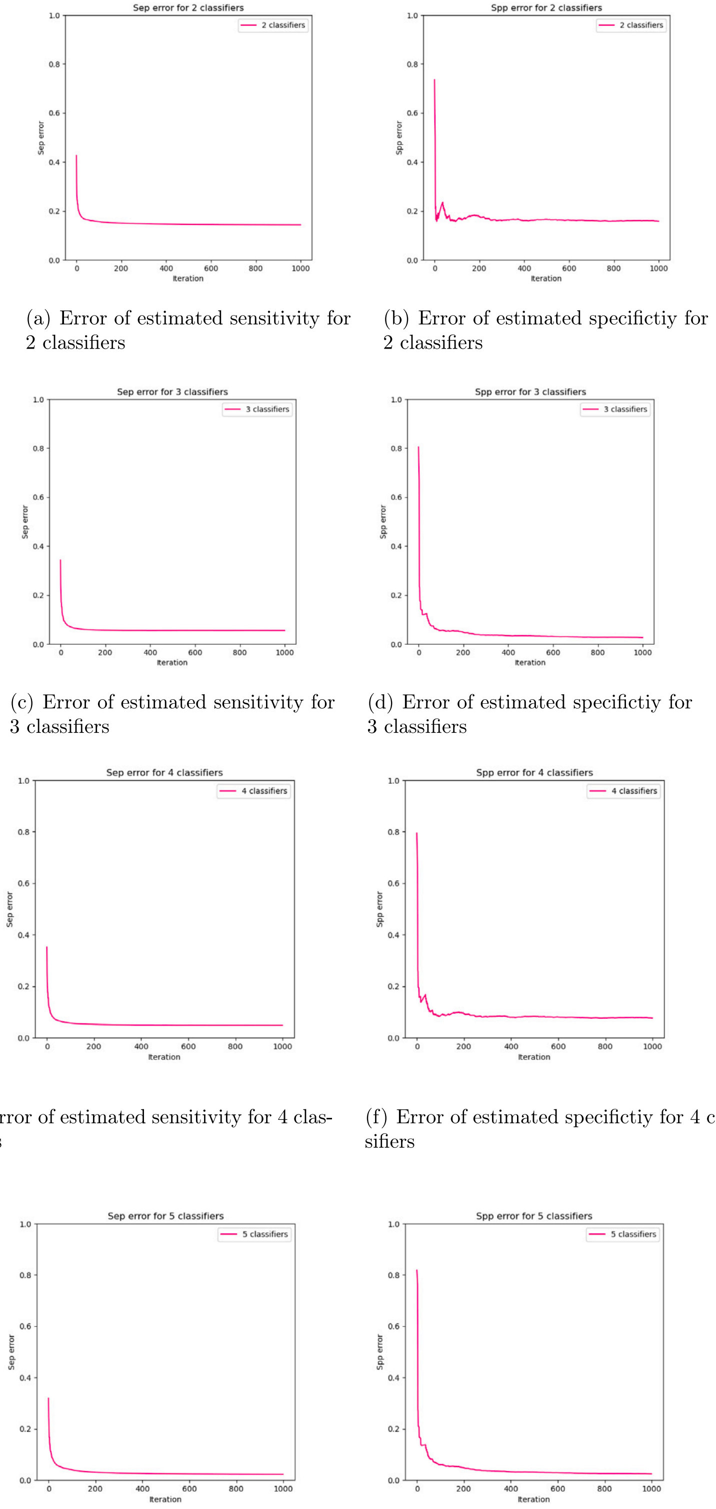
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554, and from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 826078 (Feature Cloud), and No. 101079183 (BioMedAI TWINNING). This publication reflects only the authors’ view and the European Commission is not responsible for any use that may be made of the information it contains. NDR is supported by the Knowledge Generation Project PID2023-149128NB-I00, funded by the Ministry of Science, Innovation and Universities of Spain (Proyectos de generación de conocimiento n° PID2023-149128NB-I00).

## Appendix

See Fig. 7.



(a) Error of estimated sensitivity for 2 classifiers

(b) Error of estimated specifictiy for 2 classifiers

(c) Error of estimated sensitivity for 3 classifiers

(d) Error of estimated specifictiy for 3 classifiers

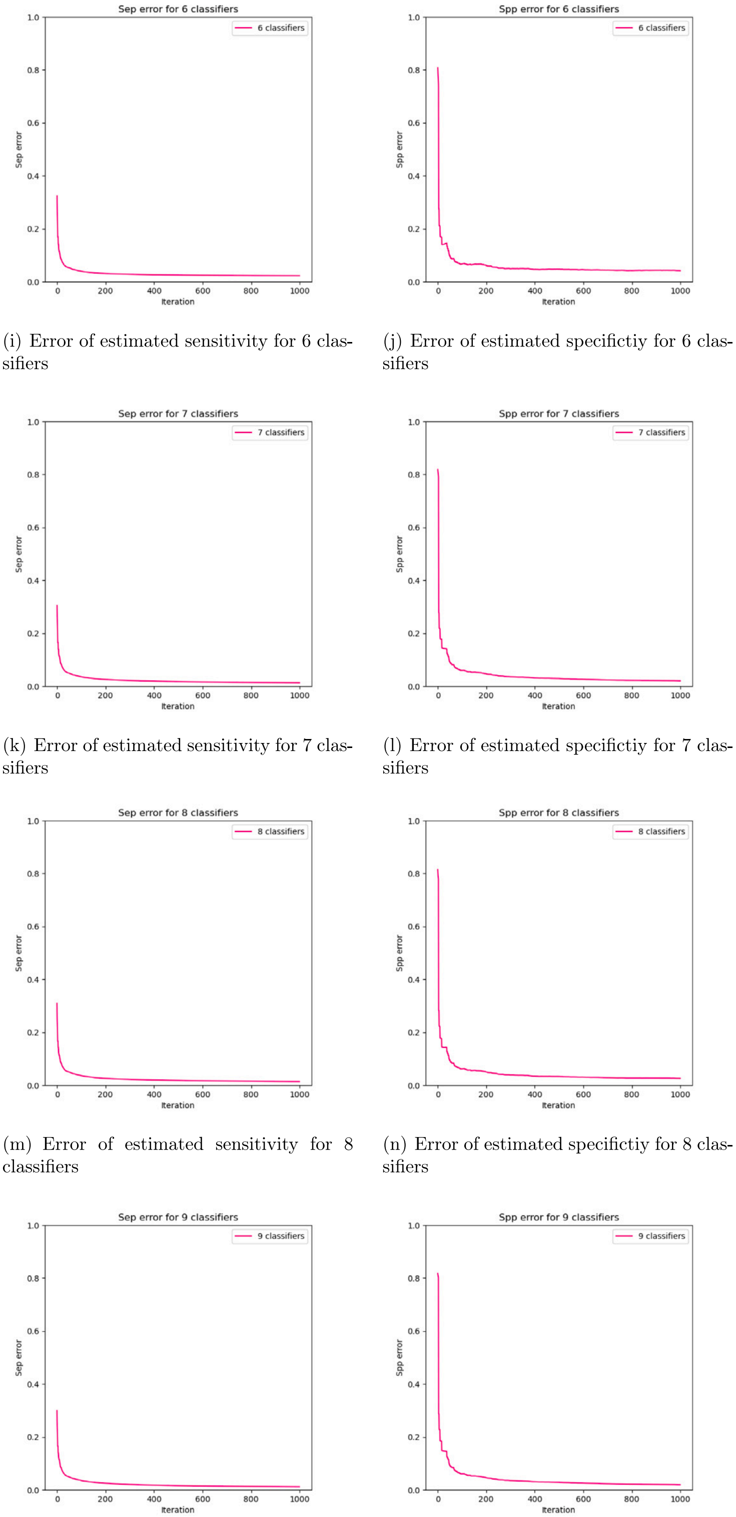
(e) Error of estimated sensitivity for 4 classifiers

(f) Error of estimated specifictiy for 4 classifiers

(g) Error of estimated sensitivity for 5 classifiers

(h) Error of estimated specifictiy for 5 classifiers

**Fig. 7.** Error of the estimation of sensitivity and specificity without known ground truth for a different number of involved classifiers.



(i) Error of estimated sensitivity for 6 classifiers

(j) Error of estimated specifictiy for 6 classifiers

(k) Error of estimated sensitivity for 7 classifiers

(l) Error of estimated specifictiy for 7 classifiers

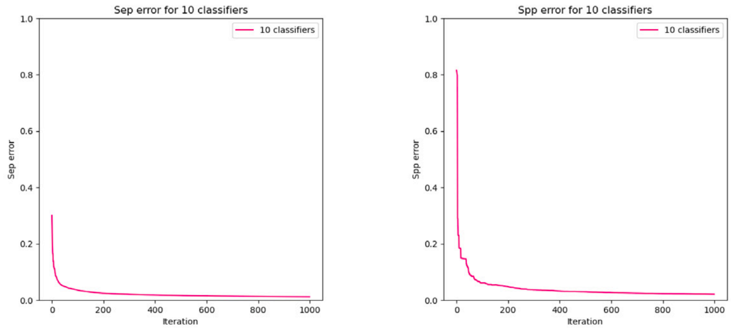
(m) Error of estimated sensitivity for 8 classifiers

(n) Error of estimated specifictiy for 8 classifiers

(o) Error of estimated sensitivity for 9 classifiers

(p) Error of estimated specifictiy for 9 classifiers

**Fig. 7.** (*continued*).



(q) Error of estimated sensitivity for 10 classifiers

(r) Error of estimated specifictiy for 10 classifiers

**Fig. 7.** (*continued*).

## Data availability

all data is available via our GitHub Repository, exclusively developed for this contribution.

## References

[1] N. Wognum, C. Bil, F. Elgh, M. Peruzzini, J. Stjepandić, W. Verhagen, Transdisciplinary engineering research challenges, in: M. Peruzzini, M. Pellicciari, C. Bil, J. Stjepandić, N. Wognum (Eds.), Transdisciplinary Engineering Methods for Social Innovation of Industry 4.0, IOS Press, Modena (Italy), 2018, pp. 753–762, http://dx.doi.org/10.3233/978-1-61499-898-3-753.

[2] N. Wognum, C. Bil, F. Elgh, M. Peruzzini, J. Stjepandić, W.J. Verhagen, Transdisciplinary systems engineering: implications, challenges and research agenda, Int. J. Agil. Syst. Manag. 12 (1) (2019) 58–89, http://dx.doi.org/10.1504/IJASM.2019.098728.

[3] G. Fabio, C. Giuditta, P. Margherita, R. Raffaeli, A human-centric methodology for the co-evolution of operators' skills, digital tools and user interfaces to support the operator 4.0, Robot. Comput.-Integr. Manuf. 91 (2025) 102,854, http://dx.doi.org/10.1016/j.rcim.2024.102854.

[4] B. Shneiderman, Human-Centered AI, Oxford University Press, Oxford, 2022.

[5] A. Holzinger, I. Fister Jr., I. Fister, H.P. Kaul, S. Asseng, Human-centered ai in smart farming: Towards agriculture 5.0, IEEE Access 12 (2024) 62,199–62,214, http://dx.doi.org/10.1109/ACCESS.2024.3395532.

[6] A. Holzinger, J. Schweier, C. Gollob, A. Nothdurft, H. Hasenauer, T. Kirisits, C. Häggström, R. Visser, R. Cavalli, R. Spinelli, K. Stampfer, From industry 5.0 to forestry 5.0: Bridging the gap with human-centered artificial intelligence, Curr. For. Rep. 10 (6) (2024) 442–455, http://dx.doi.org/10.1007/s40725-024-00231-7.

[7] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K.Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.C. Zhu, Y. Guo, W. Gao, Ai alignment: A comprehensive survey, 2023, http://dx.doi.org/10.48550/ARXIV.2310.19852, URL https://arxiv.org/abs/2310.19852.

[8] A. Holzinger, K. Zatloukal, H. Müller, Is human oversight to AI systems still possible? New Biotechnol. 85 (2024) 59–62, http://dx.doi.org/10.1016/j.nbt.2024.12.003.

[9] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J.C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, R. Perrault, Artificial intelligence index report 2023, URL http://arxiv.org/pdf/2310.03715.pdf.

[10] K. Hu, Chatgpt sets record for fastest-growing user base - analyst note, 2023, Reuters. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. (2023) 1–54, http://dx.doi.org/10.1145/3703155.

[12] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artif. Intell. Rev. 56 (4) (2023) 3005–3054, http://dx.doi.org/10.1007/s10462-022-10246-w.

[13] C.O. Retzlaff, S. Das, C. Wayllace, P. Mousavi, M. Afshari, T. Yang, A. Saranti, A. Angerschmid, M.E. Taylor, A. Holzinger, Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities, J. Artif. Intell. Res. (JAIR) 79 (1) (2024) 349–415, http://dx.doi.org/10.1613/jair.1.15348.

[14] A. Agrawal, R. Thiel, P. Jain, V. Singh, M. Fischer, Digital twin: Where do humans fit in? Autom. Constr. 148 (2023) 104,749, http://dx.doi.org/10.1016/j.autcon.2023.104749.

[15] E. Mosqueira-Rey, E. Hernandez-Pereira, J. Bobes-Bascaran, D. Alonso-Rios, A. Perez-Sanchez, A. Fernandez-Leal, V. Moret-Bonillo, Y. Vidal-Insua, F. Vazquez-Rivera, Addressing the data bottleneck in medical deep learning models using a human-in-the-loop machine learning approach, Neural Comput. Appl. 36 (5) (2024) 2597–2616, http://dx.doi.org/10.1007/s00521-023-09197-2.

[16] A. Renda, S. Schwaag-Serger, D. Tataj, A. Morlet, D. Isaksson, F. Martins, M. Mir Roca, C. Hidalgo, A. Huang, S. Dixson-Declève, P.A. Balland, F. Bria, C. Charveriat, K. Dunlop, E. Giovannini, Industry 5.0: A transformative vision for Europe, in: Governing Systemic Transformations Towards a Sustainable Industry, Publications Office of the European Union, Luxemburg, 2022, http://dx.doi.org/10.2777/17322.

[17] C. Kraišniković, R. Harb, M. Plass, W. Al Zoughbi, A. Holzinger, H. Müller, Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making, Eng. Appl. Artif. Intell. 139 (2025) 109,561, http://dx.doi.org/10.1016/j.engappai.2024.109561.

[18] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A.T. Luu, W. Bi, F. Shi, S. Shi, Siren's song in the ai ocean: A survey on hallucination in large language models, 2023, http://dx.doi.org/10.48550/ARXIV.2309.01219.

[19] M. Peruzzini, F. Grandi, M. Pellicciari, Exploring the potential of operator 4.0 interface and monitoring, Comput. Ind. Eng. 139 (2020) 105,600, http://dx.doi.org/10.1016/j.cie.2018.12.047.

[20] D. Jelonek, T. Nitkiewicz, P. Koomsap, Soft skills of engineers in view of industry 4.0 challenges, in: Conference Quality Production Improvement – CQPI, Vol. 2, No. 1, 2020, pp. 107–116, http://dx.doi.org/10.2478/cqpi-2020-0013.

[21] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A.y. Arcas, Communication-efficient learning of deep networks from decentralized data, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, Vol. 54, PMLR, 2017, pp. 1273–1282, URL https://proceedings.mlr.press/v54/mcmahan17a.html.

[22] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. (TIST) 10 (2) (2019) 1–19, http://dx.doi.org/10.1145/3298981.

[23] M.V. Luzón, N. Rodríguez-Barroso, A. Argente-Garrido, D. Jiménez-López, J.M. Moyano, J. Del Ser, W. Ding, F. Herrera, A tutorial on federated learning from theory to practice: Foundations, software frameworks, exemplary use cases, and selected trends, IEEE/ CAA J. Autom. Sin. 11 (4) (2024) 824–850, http://dx.doi.org/10.1109/JAS.2024.124215.

[24] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R.G.L. D'Oliveira, H. Eichner, S.E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P.B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S.U. Stich, Z. Sun, A.T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F.X. Yu, H. Yu, S. Zhao, Advances and open problems in federated learning, Found. Trends Mach. Learn. 14 (1–2) (2021) 1–210, http://dx.doi.org/10.1561/2200000083.

[25] J. Matschinske, The featurecloud platform for federated learning in biomedicine: unified approach, J. Med. Internet Res. 25 (1) (2023) e42,621, http://dx.doi.org/10.2196/42621.

[26] B. Pfeifer, H. Chereda, R. Martin, A. Saranti, S. Clemens, A.C. Hauschild, T. Beißbarth, A. Holzinger, D. Heider, Ensemble-gnn: federated ensemble learning with graph neural networks for disease module discovery and classification, Bioinform. (Oxf. England) 39 (11) (2023) http://dx.doi.org/10.1093/bioinformatics/btad703.

[27] C. Hausleitner, H. Mueller, A. Holzinger, B. Pfeifer, Collaborative weighting in federated graph neural networks for disease classification with the human-in-the-loop, Nat. Sci. Rep. 14 (21839) (2024) 1–10, http://dx.doi.org/10.1038/s41598-024-72748-7.

[28] A. Costantini, G. Di Modica, J.C. Ahouangonou, D.C. Duma, B. Martelli, M. Galletti, M. Antonacci, D. Nehls, P. Bellavista, C. Delamarre, D. Cesini, Iotwins: Toward implementation of distributed digital twins in industry 4.0 settings, Computers 11 (5) (2022) http://dx.doi.org/10.3390/computers11050067.

[29] A. Holzinger, A. Saranti, A.C. Hauschild, J. Beinecke, D. Heider, R. Roettger, H. Mueller, J. Baumbach, B. Pfeifer, Human-in-the-loop integration with domain-knowledge graphs for explainable federated deep learning, in: Lecture Notes in Computer Science (LNCS), vol. 14065, Springer, 2023, pp. 45–64, http://dx.doi.org/10.1007/978-3-031-40837-3_4.

[30] A. Holzinger, H. Mueller, Toward human-ai interfaces to support explainability and causability in medical ai, IEEE Computer 54 (10) (2021) 78–86, http://dx.doi.org/10.1109/MC.2021.3092610.

[31] J.M. Metsch, A. Saranti, A. Angerschmid, B. Pfeifer, V. Klemt, A. Holzinger, A.C. Hauschild, Clarus: An interactive explainable ai platform for manual counterfactuals in graph neural networks, J. Biomed. Informatics 150 (2) (2024) 104,600, http://dx.doi.org/10.1016/j.jbi.2024.104600.

[32] E. Mosqueira-Rey, A. Perez-Sanchez, E. Hernandez-Pereira, D. Alonso-Rios, J. Bobes-Bascaran, A. Fernandez-Leal, V. Moret-Bonillo, Y. Vidal-Insua, F. Vazquez-Rivera, Human-in-the-loop machine learning for the treatment of pancreatic cancer, in: 2023 International Joint Conference on Neural Networks, IJCNN, IEEE, 2023, http://dx.doi.org/10.1109/ijcnn54540.2023.10191456.

[33] J. Bobes-Bascarán, E. Mosqueira-Rey, D. Alonso-Ríos, Improving medical data annotation including humans in the machine learning loop, in: The 4th XoveTIC Conference, Vol. 3, XoveTIC 2021, MDPI, 2021, p. 39, http://dx.doi.org/10.3390/engproc2021007039.

[34] 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 (5) (2019) 93, http://dx.doi.org/10.1145/3236009.

[35] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[36] C. Moreira, Y.L. Chou, C. Hsieh, C. Ouyang, J.a. Pereira, J. Jorge, Benchmarking instance-centric counterfactual algorithms for xai: From white box to black box, ACM Comput. Surv. (2024) http://dx.doi.org/10.1145/3672553.

[37] B.H. van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102, 470, http://dx.doi.org/10.1016/j.media.2022.102470.

[38] T. Vermeire, D. Brughmans, S. Goethals, R.M.B. De Oliveira, D. Martens, Explainable image classification with evidence counterfactual, Pattern Anal. Appl. 25 (2) (2022) 315–335, http://dx.doi.org/10.1007/s10044-021-01055-y.

[39] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, A. Holzinger, On generating trustworthy counterfactual explanations, Inf. Sci. 655 119 (2024) 898, http://dx.doi.org/10.1016/j.ins.2023.119898.

[40] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 2376–2384.

[41] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Springer, 2008, pp. 705–718, http://dx.doi.org/10.1007/978-3-540-88693-8_52.

[42] A.M. Leventi-Peetz, K. Weber, Rashomon effect and consistency in explainable artificial intelligence (xai), in: A. K. (Ed.), Proceedings of the Future Technologies Conference FTC 2022 796–808, Springer, 2023, http://dx.doi.org/10.1007/978-3-031-18461-1_52.

[43] D. Brughmans, L. Melis, D. Martens, Disagreement amongst counterfactual explanations: how transparency can be misleading, Trans. Oper. Res. (TOP) (2024) 1–34, http://dx.doi.org/10.1007/s11750-024-00670-2.

[44] D. Balg, Moral disagreement and moral education: What's the problem? Ethical Theory Moral Pr. 27 (1) (2024) 5–24, http://dx.doi.org/10.1007/s10677-023-10399-9.

[45] A. Niblett, A. Yoon, Ai and the nature of disagreement, Phil. Trans. R. Soc. A 382 (2270) (2024) 20230,162, http://dx.doi.org/10.1098/rsta.2023.0162.

[46] M. Sanchez, K. Alford, V. Krishna, T.M. Huynh, C.D. Nguyen, M.P. Lungren, S.Q. Truong, P. Rajpurkar, Ai-clinician collaboration via disagreement prediction: A decision pipeline and retrospective analysis of real-world radiologist-ai interactions, Cell Rep. Med. 4 (10) (2023) 1–12, http://dx.doi.org/10.1016/j.xcrm.2023.101207.

[47] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, H. Lakkaraju, The disagreement problem in explainable machine learning: A practitioner's perspective, 2022, http://dx.doi.org/10.48550/arXiv.2202.01602, arXiv:2202.01602.

[48] Y.L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inf. Fusion 81 (5) (2022) 59–83, http://dx.doi.org/10.1016/j.inffus.2021.11.003.

[49] G. Laberge, Y.B. Pequignot, M. Marchand, F. Khomh, Tackling the xai disagreement problem with regional explanations, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 2017–2025.

[50] N. Seliya, T.M. Khoshgoftaar, J. van Hulse, A study on the relationships of classifier performance metrics, in: 2009 21st IEEE International Conference on Tools with Artificial Intelligence, IEEE, 2009, pp. 59–66, http://dx.doi.org/10.1109/ICTAI.2009.25.

[51] R. Trevethan, Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice, Front. Public Heal. 5 (2017) 307, http://dx.doi.org/10.3389/fpubh.2017.00307.

[52] S.H. Keddie, O. Baerenbold, R.H. Keogh, J. Bradley, Estimating sensitivity and specificity of diagnostic tests using latent class models that account for conditional dependence between tests: a simulation study, BMC Med. Res. Methodol. 23 (1) (2023) 58, http://dx.doi.org/10.1186/s12874-023-01873-0.

[53] R. Tripathi, S. Jagannathan, B. Dhamodharaswamy, Estimating precisions for multiple binary classifiers under limited samples, in: Y. Dong (Ed.), Machine Learning and Knowledge Discovery in Databases : European Conference, ECML PKDD 2020, Ghent, Belgium, September (2020) 14-18 : Proceedings, Parts I-V, Springer, 2020, pp. 240–256.

[54] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, TRAC Trends Anal. Chem. 80 (2016) 612–624, http://dx.doi.org/10.1016/j.trac.2016.04.021.

[55] C. Parker, On measuring the performance of binary classifiers, Knowl. Inf. Syst. 35 (1) (2013) http://dx.doi.org/10.1007/s10115-012-0558-x.

[56] P. Helber, B. Bischke, A. Dengel, D. Borth, Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019, URL https://arxiv.org/abs/1709.00029.

[57] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, Tech. Rep. 0, University of Toronto, Toronto, Ontario, 2009, URL https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.