Neural Computing and Applications
https://doi.org/10.1007/s00521-023-09197-2

ORIGINAL ARTICLE

Addressing the data bottleneck in medical deep learning models using a human-in-the-loop machine learning approach

Eduardo Mosqueira-Rey[1] · Elena Hernández-Pereira[1] · José Bobes-Bascarán[1] · David Alonso-Ríos[1] · Alberto Pérez-Sánchez[1] · Ángel Fernández-Leal[1] · Vicente Moret-Bonillo[1] · Yolanda Vidal-Ínsua[2] · Francisca Vázquez-Rivera[2]

Received: 21 May 2023 / Accepted: 20 October 2023
© The Author(s) 2023

Abstract

Any machine learning (ML) model is highly dependent on the data it uses for learning, and this is even more important in the case of deep learning models. The problem is a data bottleneck, i.e. the difficulty in obtaining an adequate number of cases and quality data. Another issue is improving the learning process, which can be done by actively introducing experts into the learning loop, in what is known as human-in-the-loop (HITL) ML. We describe an ML model based on a neural network in which HITL techniques were used to resolve the data bottleneck problem for the treatment of pancreatic cancer. We first augmented the dataset using synthetic cases created by a generative adversarial network. We then launched an active learning (AL) process involving human experts as oracles to label both new cases and cases by the network found to be suspect. This AL process was carried out simultaneously with an interactive ML process in which feedback was obtained from humans in order to develop better synthetic cases for each iteration of training. We discuss the challenges involved in including humans in the learning process, especially in relation to human–computer interaction, which is acquiring great importance in building ML models and can condition the success of a HITL approach. This paper also discusses the methodological approach adopted to address these challenges.

Keywords Human-in-the-loop machine learning · Active learning · Interactive machine learning · Pancreatic cancer · Generative adversarial network

Mathematics Subject Classification 68T05 · 68T07

1 Introduction

1.1 Data bottleneck

Anyone who has ever developed a machine learning (ML) model is aware that the first problem they face is obtaining sufficient and representative data to be able to successfully implement training. This problem has been exacerbated in recent years with deep learning (DL) algorithms that require a vast amount of data for training.

The ML developer is thus confronted with two possibilities: collect the data oneself, or rely on public data collected by others. They each have drawbacks, the first because the process can be demanding in terms of both human and time resources and, depending on the circumstances, data may even be impossible to obtain, and the second because public data are often scattered, difficult to locate, may not correspond exactly to the problem to be solved, and often have issues that limit their applicability, such as inconsistencies, missing values, class imbalance, and so on.

This problem has come to be known as the *data bottleneck* [77], defined as the inability to locate quality data with which to train ML models. And we say *quality data* because the problem often lies not only in the number of cases available, but also in their quality, which is not always easy to measure [8].

There are several ways to deal with data bottlenecks. One is to develop open datasets, curated around unsolved problems and made available to researchers. Examples are the Nightingale Open Science initiative [58] for the field of medicine in general, and The Cancer Genome Atlas

Extended author information available on the last page of the article

Program [83], which makes a large number of diagnosed cancer cases with all the related data available to researchers.

But even with these initiatives, it is very likely that the datasets used have certain issues that may affect the performance of the ML models developed using them. Techniques to reduce the impact of these issues have been organized into two main groups [77]:

- *Data are missing.* If data are missing, we can create more data using techniques such as *data augmentation*, or get more out of existing data with techniques such as *curriculum learning*, or reuse a model trained with other data as a starting point for our problem using techniques such as *transfer learning*.
- *Labels are missing.* If labels are missing, techniques such as *active learning* or *gamification* can be used to create them, or we can create what are called *weak labels* using heuristically generated label functions and external knowledge bases to programmatically label the data.

1.2 Human-in-the-loop machine learning

We need to take into account not only the data we are going to use in ML, but also the learning process itself. Until recently, ML models were built by humans going through steps as follows: obtaining data, preprocessing the data, performing feature engineering, launching learning, and tuning learning hyperparameters trying to improve the results. However, an important paradigm shift occurred with the advent of DL models [44]. In these models, feature extraction is algorithmically computed, without human intervention, using a series of layers that, starting with the raw input, transform a representation at a lower level into a representation at a higher and slightly more abstract level.

Recently, techniques have been developed that include human participation in the learning process (in some aspect related to it). These techniques are often collectively referred to as human-in-the-loop (HITL) ML [56].

DL has been largely based on taking humans out of the equation to improve performance; however, this has come at the cost of needing more data and greater computational requirements. The idea behind HITL-ML is to overcome those obstacles: to make learning more efficient by using fewer data and fewer computational resources. This is especially true in medical domains, where human expertise and extensive experience can fill in the gaps in large amounts of data or help deal with complex data [29].

We can classify HITL methods according to their relationship to the learning process [56], as depicted in Fig. 1.

- *Before learning.* Here, we find *curriculum learning (CL)* [10], in which the dataset is organized in terms of increasing complexity in order to take advantage of previously learned concepts and to ease the abstraction of new concepts. This ordering can be done automatically or by domain experts.
- *During learning.* Based on the entity in control of the learning process, three distinct categories of techniques can be identified [28]: (a) *active learning (AL)* [76], where the system controls the learning process and relies on human input to label unlabelled data; (b) *interactive ML (IML)* [7], characterized by closer interaction between users and learning systems, where humans frequently and interactively provide information to the system and react to the system's responses; and (c) *machine teaching (MT)* [70, 79], where human domain experts use didactic techniques to control the learning process without needing any ML skills or expertise.
- *After learning.* DL models are characterized by being poorly explainable since they lack a declarative representation of knowledge that is interpretable [30]. *Explainable artificial intelligence (XAI)* [22] tries to produce more explainable models while maintaining a high level of learning performance.
- *Beyond learning.* Finally, we can include techniques that are beyond the learning process, such as those involved in *usable AI* and *useful AI* [90]. The former refers to AI solutions that ensure an optimal user experience, and the latter, going beyond usability, refers to how to develop ML solutions that satisfy user needs in a trustworthy manner.

In the recent literature, the term HITL-ML is mainly associated with AL, as is demonstrated in Chen et al. [14], Na et al. [60], and Saghir et al. [75], although works also exist that cannot be included in the previous categories, such as Delussu et al. [16], in which user feedback is used to identify people in images in the domain shift environment. Also, in Abdar et al. [1], a novel ensemble learning approach is proposed that includes late fusion in both feature selection and decision steps. The novelty lies in using feature selection by both machine and human experts and then applying the ensemble technique.

HITL can be used also not only to obtain better performance in ML models, but to achieve a new type of relationship between humans and these models. For example, Mosqueira-Rey et al. [55] used MT based on didactic techniques as a didactic technique itself to teach orthography to students.

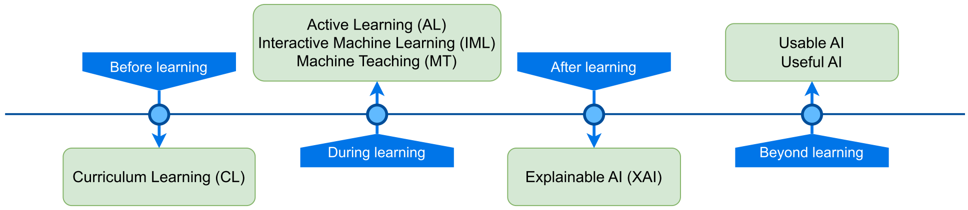


Fig. 1 Classification of HITL methods

1.3 Contribution

The threefold contribution of this paper is as follows. First, we demonstrate that involving human experts in the learning process improves the learning capacity of a neural network model. This is especially important in the medical domain where data are usually difficult to obtain. Second, we present specific HITL strategies to address the *data bottleneck* problem, whether a *data missing* problem or a *labels missing* problem.

The strategy followed to solve the *missing data* problem is a data augmentation process carried out through a conditional tabular generative adversarial network (CTGAN). The novelty of this approach, further explained in Sect. 5, is that humans, acting as an additional discrimination layer in the GAN, try to identify synthetic cases and, through an IML process, provide the information that allows identification of those cases as not real. That information is subsequently converted into a new *condition* or *constraint* that is applied to the next synthetic cases generated by the CTGAN, so that, in successive iterations, cases are more indistinguishable from real cases (and therefore more useful for learning).

That leaves us with a sort of *labels missing* problem, and we say "sort of" because all the cases have labels, but we can consider them to be weak for two reasons: first, the labels of the synthetic cases have been assigned by the ML model itself, so we cannot consider them to be entirely reliable; and second, the actual labels of the dataset are also unreliable, given that several valid courses of action are possible in a complex medical environment, based on protocols that may change over time. So here an AL strategy is followed, whereby we do not relabel the entire dataset, as this would be an unrealistic goal, but only those cases that the model considers doubtful (see Sect. 4).

Finally, while bringing humans back into the loop in ML offers advantages, it also implies a new set of very human problems such as availability, attention, interactivity, and different expertise. Our contribution to addressing these human issues in the HITL approach is a usability analysis of the whole process of interaction between the experts and the model using an extended usability model and a context-of-use taxonomy (see Sect. 6). The idea is to ensure that expert interactions with the system are simple and do not imply a high cognitive load beyond that inherent to the complex problem they are dealing with. The experts can thus focus their attention more on the problem to be solved and less on the application used to present the cases to them.

The paper is structured as follows: in Sect. 2, we describe a pancreatic cancer problem and the corresponding dataset. In Sect. 3, we provide a general overview of the experiment and describe the artificial neural network (ANN) used. In Sect. 4, we explain the AL approach in more detail, and in Sect. 5, we do the same for the IML approach. In Sect. 6, we describe the human–computer interface (HCI) issues that we faced and how we solved them. Finally, we report our results in Sect. 7 and include a discussion, final conclusions, and pointers for future work in Sect. 8.

2 Pancreatic cancer

Pancreatic cancer incidence and mortality are both high and symptoms are frequently absent. Crucial in diagnosis is correct identification of pancreatic tumours, which can be classified as [53]: (a) neoplasms of the exocrine pancreas, (b) neoplasms of mixed or uncertain differentiation, (c) tumours of the endocrine pancreas, (d) pancreatic mesenchymal tumours, and (e) secondary tumours of the pancreas.

Of these tumour types, we focus on pancreatic adenocarcinomas, a type of neoplasm of the exocrine pancreas, because of their higher incidence. We used as reference "*Pancreatic Adenocarcinoma—NCCN Clinical Practice Guidelines in Oncology*" [61], which is widely used by the medical community for this type of diagnosis.

Numerous papers describe applications of AL and IML to the diagnosis of pancreatic cancer, although since they mainly focus on tumour image analysis, they cannot be used as a reference for a guideline-based analysis. For example, Wen et al. [88], in a study of the application of AL to segmentation quality assessment of pancreatic cancer images, reported satisfactory performance and efficiency for three classification methods, namely, support vector machine (SVM), random forest (RF), and convolutional neural network (CNN). Another noteworthy application was by Zhuang [94], who applied DL techniques to the interpretation of computed tomography images of pancreatic lesions, and pancreatic neuroendocrine tumours.

In our work, we applied AL and IML techniques to pancreatic cancer diagnostics, based on variables used and tests performed by physicians in accordance with clinical practice guidelines in oncology, with the aim of developing an approach to AL and IML diagnostics that is close to the diagnostic process usually followed by medical staff.

The dataset used in this work was obtained from The Cancer Genome Atlas Program [83]—published by the USA National Cancer Institute (NCI) and the National Human Genome Research Institute—as the database of pancreatic cancer cases most widely used in this type of study. This database is composed of several research projects, among them, TCGA-PAAD, currently with 185 diagnosed cases with all the necessary details to carry out a full analysis of pancreatic cancer cases, including their treatments.

The TCGA-PAAD project consists of information about cancer patients. The raw data have a total number of 158 attributes, but since some of them were irrelevant to the problem in hand (that is the chemotherapy treatment decision), some of them were not used (e.g. the project code, the disease code, the therapy ongoing, etc.). Finally, only 56 attributes were considered. Preprocessing included the removal of duplicate cases, the removal of irrelevant columns, and refactoring of the data labels. The data for 185 patients (83 female and 102 male) indicated that they were cancer positive for three disease types: adenomas and adenocarcinomas, ductal and lobular neoplasms, and cystic, mucinous, and serous neoplasms. For each case, we were interested in determining whether chemotherapy treatment was indicated or not based on the diagnostic information available. The database includes patient demographic information, family history, diagnosis, treatments, and genomic, epigenomic, transcriptomic, and proteomic data.

Other scientific studies have used the same dataset for different purposes, e.g. separation of cases into moderate and aggressive clusters in order to develop a prognosis and survival rate model, also using the genetic information provided in the dataset [39], development of a DL model to identify pancreatic cancer subtypes and determine their molecular characteristics [80], and data curation to identify biomarkers [62].

3 Experiment design

The experiment took place over one month. On weekdays, the patient samples were assessed and annotated by a panel of cancer experts (two to four, depending on availability), and at weekends, the system was retrained with the newly annotated data.

The workflow of the HITL system is shown in Fig. 2 and its steps were as follows:

1. *Training the ANN model with the initial dataset.* This model was the baseline from which we started training (ANN characteristics are explained in more detail in Sect. 3.1).
2. *Applying data augmentation using a CTGAN.* A CTGAN was trained to generate synthetic cases that would augment the dataset (CTGANs are described in more detail in Sect. 5.1).
3. *Making predictions using the model.* The model predicted the labels of the real and synthetic data.
4. *AL—uncertainty sampling.* The human expert was provided with mostly synthetic cases for labelling, but also some real cases. Here, the system followed an uncertainty sampling strategy to select cases close to the decision boundary, i.e. those that have the highest uncertainty in the classification (details of the AL experiment are described in Sect. 4).
5. *Including cases considered certain in the dataset.* Cases with predictions that were considered certain were sent to the dataset, including synthetic cases not selected for labelling by the expert and whose labels were determined by the model's predictions.
6. *AL—new data labelling.* The human experts labelled the data presented to them and these data were also added to the dataset. Therefore, the new dataset included both cases for which predictions by the model were certain and cases reviewed by human experts.
7. *IML—CTGAN constraint updating*. The experts identified inconsistencies in synthetic cases that led to their rejection. Those inconsistencies were used to create new constraints that would feed into creating better synthetic cases in subsequent iterations (CTGAN updating is described in more detail in Sect. 5.2).
8. *Re-training the model.* The model was retrained with the new dataset.
9. *Re-training the CTGAN.* The CTGAN was retrained if new constraints were included in the IML experiment.

For each case presented to the experts, we collected the following information:

- *Patient treatment.* Whether or not the patient should receive chemotherapy given the data available.
- *Reason for prescribing chemotherapy.* In an area as sensitive as health care, understanding the reason for indicating chemotherapy.
- *Identification of synthetic cases.* Whether or not the expert considers a given case to be synthetic.

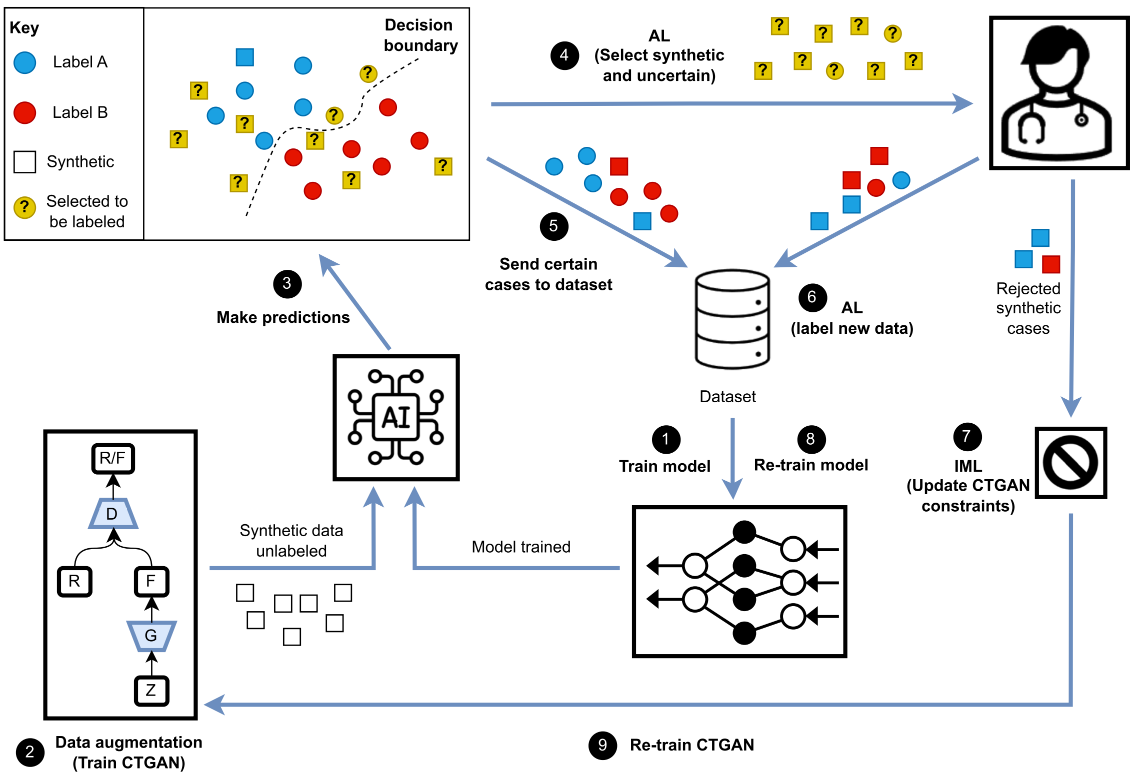


Fig. 2 HITL experiment workflow

- *Comments on cases.* Feedback as shared by the experts, which usually included comments regarding the data, possible failures detected in the data and, mainly, an explanation of why case was identified as synthetic.

The interface used by the experts in the AL and IML processes, as well as the related HCI aspects, is detailed in Sect. 6.

3.1 ML model: ANN

An ANN was used as the ML model, preferred over other "shallow" ML models because an ANN gives us the flexibility to scale up the problem to more complex domains if necessary [74]. Since the dataset consisted of structured data, it was decided to use a simple dense network instead of more complex versions such as convolutional networks or recurrent networks, more suitable for handling unstructured data or numerical series.

The inputs of the model were the 56 features that described each patient. The output of the last layer was associated with the classification problem possible categories, i.e. the options of "Chemotherapy" or "No Chemotherapy".

The hyperparameters of the ANN (number and size of hidden layers, learning rate, momentum, batch size, etc.) where established by an optimization process called *grid search*, consisting of an exhaustive search of the hyperparameters in a range of values provided by the ML engineer. While grid search, because it tests all possible configurations of hyperparameters, is a very computationally expensive technique, it is the preferred solution in low-dimensional spaces because of ease of execution, parallelization, and durability [9]. The values used for each hyperparameter are summarized in Table 1.

Because of the small size of the dataset, regularization was necessary to avoid overfitting (the model fitting the training data so well that it loses its ability to generalize and so predictions for new cases are incorrect). Therefore, we included the following regularization techniques: (1)

Table 1 Range of the hyperparameter values

Hyperparameter	Values
Hidden layers	1, 2
Neurons in each layer	64, 128, 256
Learning rate	$1e^{-2}, 1e^{-3}, 1e^{-4}$
Momentum	0.95, 0.90, 0.85, 0.80
Dropout	0.3, 0.4
Batch size	16, 32, 64
Epochs	10, 20, 30, 40

dropout layers were added to each hidden layer to eliminate co-adaptation between neurons, (2) an L2 regularization term was added to restrict the values of the weights to small numbers, and (3) 10-fold cross-validation was performed to avoid overdependence on the data selected as the validation dataset. The final model was an ANN with two hidden layers with 128 neurons in each, a learning rate of 0.9, a momentum of 0.9, a dropout value of 0.4, a batch size of 16 and 20 epochs. Figure 3 shows the layer distribution of the *base model*.

4 AL approach

AL is a machine learning technique used to overcome labelling bottleneck by posing queries in the form of unlabelled instances to be labelled by an oracle (e.g. a human annotator) [76].

The goal of AL is to use fewer training examples than other ML techniques to achieve the same accuracy. It is particularly useful when the labelling process is expensive or time-consuming, or when dealing with a scenario of scarcity of examples.

AL is also useful in weak supervision scenarios [12, 47], in which labelling functions that encode domain knowledge—such as user-specified heuristics or external knowledge bases—are developed and used to noisily annotate subsets of data. The weak labels generated in this process may not be very reliable, so human supervision is necessary to relabel those cases identified as doubtful by the model.

In our case, the *weak labels* came from the synthetic cases generated by the CTGAN and annotated by the current iteration of the model. As the cases were generated and the labels were created by a model that was not fully trained, we could not have too much confidence in them. We could also have doubts regarding the existing labels in the dataset itself, not so much because they were erroneously labelled, but because they originated with physicians who applied clinical practice guidelines that were changed and updated. Therefore, it was necessary to analyse and relabel them according to more current guidelines.

Regarding the AL approach employed, an essential stage in any AL procedure involves defining the sampling process, also known as the query strategy, which entails the selection of instances to be labelled by the human expert. Two options exist [59]:

- *Uncertainty sampling* identifies unlabelled items that are close to a decision boundary in the current ML model.
- *Diversity sampling* identifies unlabelled items that are underrepresented in or unknown to the current ML model.

These two types of sampling correspond to a well-known dilemma in AI: *exploitation* versus *exploration* [26]. Uncertainty sampling is an exploitation process in which the focus is on improving efficiency using existing data, whereas diversity sampling is an exploratory process that tries to go beyond the known data samples to enhance the diversity of the data.

In our case we decided initially to apply uncertainty sampling, mainly due to the size of the dataset: since we had few cases, selecting cases with the highest uncertainty was more important than diversity. However, our sampling process had the peculiarity that our major source of uncertainty was the synthetic cases generated by the CTGAN, so we ultimately followed a mixed strategy: 80% of the cases selected for labelling by the expert came from the CTGAN and the remaining 20% were cases near the decision boundary (synthetic or real).

Lastly, another aspect associated with the AL process is determining the number of new instances to label before retraining the model. In this context, Rubens et al. [73] identified two primary approaches:

- *Batch.* Multiple examples are labelled before the model is retrained, with the batch size also impacting on the model's performance.
- *Sequential.* The system undergoes retraining after each new labelling of elements, providing immediate feedback to the user.

When considering these alternatives, various trade-offs arise. Sequential training is crucial in recommender systems, as users expect to receive an updated list of recommendations based on their latest annotation. Small batch

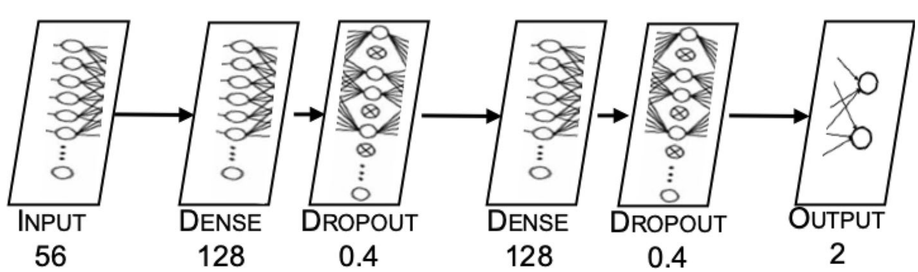


Fig. 3 Layer distribution of the base model

sizes ensure that the most benefit is gained from each data point in each iteration. But these are the least efficient strategies in terms of computational cost, since the model has to be trained more often. Maximizing the sample size, however, will ensure that more items are labelled sooner and the model has to be retrained less often, making the overall process more computationally efficient.

Another cost, apart from the computational cost, that it is not usually considered is the cost associated with human interactions. One problem associated with AL is the assumption that the oracle is "*infallible (never wrong), indefatigable (always answers), individual (only one oracle), and insensitive to costs (always free or always charges the same)*" [17]. Obviously, however, this is not true, as humans can become distracted and fatigued over time, and this introduces variability in the quality of their annotations.

We consider that the best strategy to avoid including noisy annotations in the AL process is to take HCI issues into account. It is important that the number of cases in each iteration and the number of iterations are not too high. The idea is to avoid fatigue, boredom, loss of interest, etc., and also not to overburden professionals whose time is scarce. We therefore decided to limit batch size to ten cases and to limit the duration of the experiment to one month. The HCI issues are described in more depth in Sect. 6.

Our AL approach followed an iterative process, starting from a set of labelled (i.e. known) examples, and a set of unlabelled examples (i.e. unknown) that could be incorporated in the model. The aim is to select the best examples to both improve the learning process and to make human participation more efficient by reducing the number of examples to be annotated.

AL has been applied in several cancer diagnosis scenarios. Wen et al. [88] uses AL for segmentation quality assessment for pathology images, comparing three classification methods for performance improvement and efficiency. Halder and Kumar [25] described an AL approach that deployed a rough fuzzy classifier for cancer prediction, using micro-array gene expression data as the basis and providing an alternative to other AL algorithms. AL has also been used in a narrowing uncertainty process in two breast cancer classification experiments [45].

5 IML approach

As we have seen, in AL interactivity is limited to humans acting as annotators of the cases presented to them. But more important than the limited interactivity is the lack of control over the process, as it is the model that decides which case should be presented to the experts for annotation. IML includes a wide range of applications where control is shared between humans and the model and where interaction between them is close. This last point is important because, given the increased level of interaction in IML, it becomes necessary to consider HCI techniques.

The most typical form of IML is that the human reviews and corrects annotations made by the model on unstructured data. For example, a human can use an IML process to perform image segmentation or to fine-tune segmentation as performed by the model in a process known as interactive image segmentation [68]. Other successful applications involve working with video [40] and sound [18] time-series data.

IML, since it is based on AL, shares some of its limitations while also introducing its own. A prominent issue is the blending of ML and HCI aspects due to increased interactivity, leading to the need for more extensive efforts for application development, as these must be tailored and studied individually. Perhaps a future solution lies in exploring methodologies and theoretical frameworks for IML systems, like that proposed by Meza Martínez et al. [50].

In our case we used IML as an aid to the data augmentation process. As an additional discriminatory layer in the CTGAN in charge of generating the synthetic data, we used the experts, who were in charge of creating new constraints that would improve the generation of synthetic cases in each iteration. Below we describe the CTGAN in more detail, and also the process by which expert opinions were collected to incorporate new constraints into the system.

5.1 CTGANs

When employing supervised ML algorithms in the medical domain, one of the main challenges is to deal with small datasets and numbers of annotated samples, given that the algorithms require labelled data and a sufficient number of training examples. Researchers attempt to overcome this challenge by using data augmentation schemes, one of the most popular of which is the generative adversarial network (GAN).

A GAN [20] is a DL framework composed of two networks—a generator and a discriminator—competing with each other in the form of a zero sum game. The generator produces data examples taking into account the characteristics of the training data, and the discriminator tries to distinguish real data from generated data.

GAN models have been successfully applied in the fields of computer vision [87], natural language processing [84], and image generation [35], among others. Given their excellent performance, they have also attracted the attention of researchers in the medical image fusion field, as exemplified by Fu et al. [19], Zhan et al. [91], Jiang et al. [38], and Guo et al. [23].

Particularly useful in the data augmentation context is to allow for controlled image generation [51]. GANs conditioned on a label or a segmentation map, for instance, can be used to generate synthetic lesions or, more generally, to balance a dataset by augmenting underrepresented groups [78]. Image translation architectures, such as CycleGANs [93], have been used for cross-domain medical image synthesis, which allows samples to be transferred from modalities in which data are relatively abundant (e.g. computed tomography) to more costly or less widely implemented modalities (e.g. magnetic resonance imaging) [37].

Although the most popular studies related to GANs involve datasets from the computer vision domain, data science applications, even in the medical domain, usually deal with multiple continuous and categorical variables. Over the past 6 years, the promise of GAN models has encouraged their development for tabular data generation. However, the generation of synthetic data in tabular datasets is not so simple, as we normally have a mixture of continuous data which may have multiple modes, and discrete data which is sometimes imbalanced. Several approaches have been proposed for synthetic tabular data generation. Based on input real patient records, medGAN [15] generates high-dimensional discrete variables via a combination of an autoencoder and GANs. Mottini et al. [57] used Cramér GANs with a generator architecture that combines feed-forward layers with the Cross-Net architecture and uses an input embedding layer for the categorical features to generate realistic synthetic passenger name records. In the field of data synthesis and with the objective of maintaining privacy, Park et al. [66] proposed a method called table-GAN that synthesizes tables containing categorical, discrete, and continuous values.

GANs can be extended to a conditional model—called CGAN [51]—if both the generator and discriminator are conditioned on some extra information, which could be any kind of auxiliary information, such as class labels or data from other modalities. Conditioning is done by feeding this extra information into both the discriminator and generator as an additional input layer. CGANs have been successfully applied in many fields, including images [49], natural language [82], and anomaly detection [85].

The tabular version of the CGAN is called conditional tabular GAN (CTGAN) [89]. CTGAN allows conditions or constraints to be assigned to the synthetically generated tabular data, thus allowing values to be assigned in a fixed way or to be calculated with respect to other columns (features). These conditions improve the accuracy of the data by prohibiting combinations of feature values that may not exist in the real dataset. This very common scenario when working with tabular data is finding features that have very particular relationships between them that are very hard to model and that can easily confuse a case generator.

The CTGAN approach has been applied from different perspectives. From the perspective of data balancing, Wang et al. [86] proposed the CTT (traffic) GAN scheme to expand small category samples in traffic datasets for classification purposes. Jia et al. [36] used CTGANs to augment disk failure data, demonstrating their effectiveness through classic ML models. Nugraha et al. [65] developed a classification system to predict health insurance fraud, solving the imbalanced data problem by using CTGAN as an oversampling method to generate additional data for minority classes. In the missing data imputation field, Khan et al. [41] used CTGAN to add synthetic samples and increase the amount of training data and, in this way, improve imputation performance. From the perspective of dataset generation for domains with no public datasets, Rahman et al. [69] applied CTGAN to obtain a dataset with personality trait scores and responses to phishing with a view to investigating the psychological aspects that may contribute to sensitivity to phishing attacks. With similar aims, Tang et al. [81] augmented training data to build an ensemble ML framework to search for sweet spots in shale reservoirs, Hong and Baik [32] generated voluminous training data to establish bankruptcy predictions, and Moon et al. [52] generated electric load data to train forecasting models.

5.2 CTGAN updating using IML

We used the CTGAN implementation that is part of the Synthetic Data Vault (SDV) project [67], as it allows special relationships to be defined between columns called *constraints* that are used to improve the quality of the generated data by prohibiting certain combinations that may not exist in real data.

CTGAN allows several types of constraints. In our case we mainly used three types:

- *Fixed combinations* force combinations between a set of columns to be fixed, i.e. no other permutation or shuffling is allowed other than what is already observed in the real data. An example would be different columns with data representing cities and countries, which we would not want to be shuffled, ending up with incorrect associations such as Paris–Italy or Rome–Spain.
- *Inequalities* force inequality relationships between pairs of columns. For every row, the value in one column must be greater than the value in another column. For example, an employment start date in a company must be earlier than the employment end date.

- *Custom constraints* are used to represent business logic that cannot be represented using predefined constraints such as fixed combinations and inequalities.

Constraints can be handled in two ways:

- *Rejection.* A sample is discarded if the generated synthetic sample violates a constraint.
- *Transformation.* The data are transformed in such a way as to guarantee that the synthetic data look like the original data, e.g. by copying one of the possible combinations in the original data that meets the imposed constraints.

Although rejection is a simple procedure, it may slow down the sampling process, whereas transformation is more efficient but cannot always be used.

The *missing data* problem was solved by a data augmentation process in a CTGAN model where humans acted as additional discrimination layers. Through IML, the information provided by the experts to identify synthetic cases was converted into new *conditions* or *constraints* that improved the CTGAN results.

Figure 4 depicts a representation of several of these restrictions:

- *Constraint* `lymph_node` is an inequality constraint that was manually created as the result of an expert comment that *There cannot be more positive lymph nodes than nodes tested*, an obvious situation but one that the CTGAN failed to identify. The handling strategy is *reject*.
- *Constraint* `fixed_R1_postoperative` is a fixed combination constraint that resulted from an expert commentary that *You cannot have post-operative RXTX if you do not have radiation therapy prescribed because of a residual tumour*. The handling strategy is *reject*.
- *Constraint* `fixed_pathologic` was created because the pathological stage of patients follows rules that cause several columns to affect each other. As the number of possible combinations between the columns defining the pathological stage was very large, the CTGAN could not infer and reproduce all the relationships. Because of this, the expert detected inconsistencies such as *The stage is not correct, if there are positive nodes in He, they should also be identified in IHQ and also the stage cannot be N0*. The handling strategy is *transform*, i.e. modifying the data with some of the existing combinations in the original dataset.
- *Constraint* `days_to_new_event` is a custom constraint used when there are no new tumours after initial treatment, and so it eliminates the values of columns associated with the treatment of new tumours. The constraint identifies the affected columns and, after generating the synthetic cases, the `reverse_transform` function modifies the values of those columns to `NaN` if there are no new tumours, i.e. the handling strategy is *transform*. This constraint was added and due to numerous comments such as *It is inconsistent to have a progressive disease with the absence of further events* and *Pharmaceutical therapy is YES. Why are you treating him if the disease has not recurred?*

Our CTGAN model had a total of 19 constraints, which were included in the model as the outcome of feedback from human experts. Sixteen were fixed combination constraints, meaning that the values of the columns involved should be present in the real data, while the remaining constraints were two inequalities and one custom constraint. The main handling strategy was rejection (fourteen cases), since it is often easier to discard a conflicted case and generate a new case, and the handling strategy for the remaining cases was transformation. This IML process helped us to build better synthetic cases that were more indistinguishable from real cases and so were more useful for the training process.

6 HCI issues

Given the human-centred nature of HITL-ML, it is logical that the literature and the standards produced in the HCI field become more crucial than for ML in general. The central concept of HCI is *usability*, which is defined in the ISO 9241-210:2010 [33] standard as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*". This means that usability is not an inherent property of a system but rather depends on the characteristics of its *context of use*, where the main attributes are *users*, *tasks*, *equipment*, and *environment* (again according to ISO 9241-210:2010 [33]).

We also need to know the specific usability criteria to focus on, because, as shown by Gray and Salzman [21], a general idea that effectiveness, efficiency, and satisfaction are important is not very useful if we specifically want to identify actual usability problems. This is why some usability experts have created what Lewis [46] calls *expanded models of usability*, which consist of multiple usability attributes and subattributes, organized into some kind of hierarchy or taxonomy.

Since usability is so complex and can be assessed from so many different points of view, the usability studies we find in the real world take many forms and involve a great diversity of methods. Adelman and Riedel [2] identified 13 methods, which they classified into three types, namely: *expert* (determine what is good and bad about the system

```python
# Constraint "lymph_node"
lymph_node = GreaterThan(
    low = 'number_of_lymphnodes_positive_by_he',
    high = 'lymph_node_examined_count',
    handling_strategy = 'reject_sampling'
)

# Constraint "fixed_R1_postoperative"
fixed_R1_postoperative = FixedCombinations(
    column_names=['residual_tumor', 'radiation_therapy', 'postoperative_rx_tx'],
    handling_strategy='reject_sampling'
)

# Constraint "fixed_pathologic"
fixed_pathologic = FixedCombinations(
    column_names=['pathologic_stage', 'pathologic_T','pathologic_N', 'pathologic_M', 'residual_tumor'],
    handling_strategy='transform'
)

# Constraint "days_to_new_event"
def transform(table_data, column):
    return table_data

def reverse_transform(table_data, column):
    not_new_event = table_data.new_tumor_event_after_initial_treatment == "NO"
                    or table_data.new_tumor_event_after_initial_treatment.isna()
    table_data[column].loc[not_new_event] = np.nan
    print(table_data[column])
    return table_data

days_to_new_event = CustomConstraint(
    columns=['days_to_new_tumor_event_after_initial_treatment','new_neoplasm_event_type',
             'new_neoplasm_event_occurrence_anatomic_site','new_neoplasm_occurrence_anatomic_site_text',
             'progression_determined_by','new_tumor_event_additional_surgery_procedure',
             'days_to_new_tumor_event_additional_surgery_procedure',
             'residual_disease_post_new_tumor_event_margin_status','additional_radiation_therapy',
             'additional_pharmaceutical_therapy'],
    transform = transform,
    reverse_transform = reverse_transform
)
```

Fig. 4 CTGAN constraints obtained from feedback from human experts

from a usability perspective), *subjective* (obtain user opinions about the usability of evolving prototypes and operational systems), and *empirical* (obtain objective data about how well people can actually use a system). Ivory et al. [34] proposed a different and even more complete taxonomy of usability evaluation methods, classified into five types, namely: *testing* (users perform tasks), *inspection* (evaluators identify problems), *inquiry* (users provide feedback), *analytical modelling* (models are used to make predictions), and *simulation* (models are used to mimic interactions).

6.1 Scope of the analysis

As part of our experiment, we analysed both the context of use and the usability of the application itself. In choosing a usability taxonomy, we normally confront the issue that the concept of usability has been very inconsistently described in the literature. To provide an objective, comprehensive, and structured taxonomy, we selected as the basis for our task the expanded usability model by Alonso-Ríos et al. [4], a work that also concerns the context of use in a separate taxonomy [5]. These taxonomies have been previously used as the basis of a systematic and generalizable methodology for usability evaluation [6].

Our first step was to establish the scope of the analysed system. As mentioned previously, the ML experiment consisted of an AL process, with a first model built, without the intervention of the domain experts participating in the experiment, using the pancreatic cancer data from the dataset by means of the ANN described in Sect. 3.1. This model was then retrained with new information produced by combining existing data with new synthetic cases produced by the CTGAN, all relabelled in the AL experiment.

A web application serving as the front-end (Fig. 5) presented the cases selected by the AL sampling strategy to the human experts (i.e. medical doctors specializing in

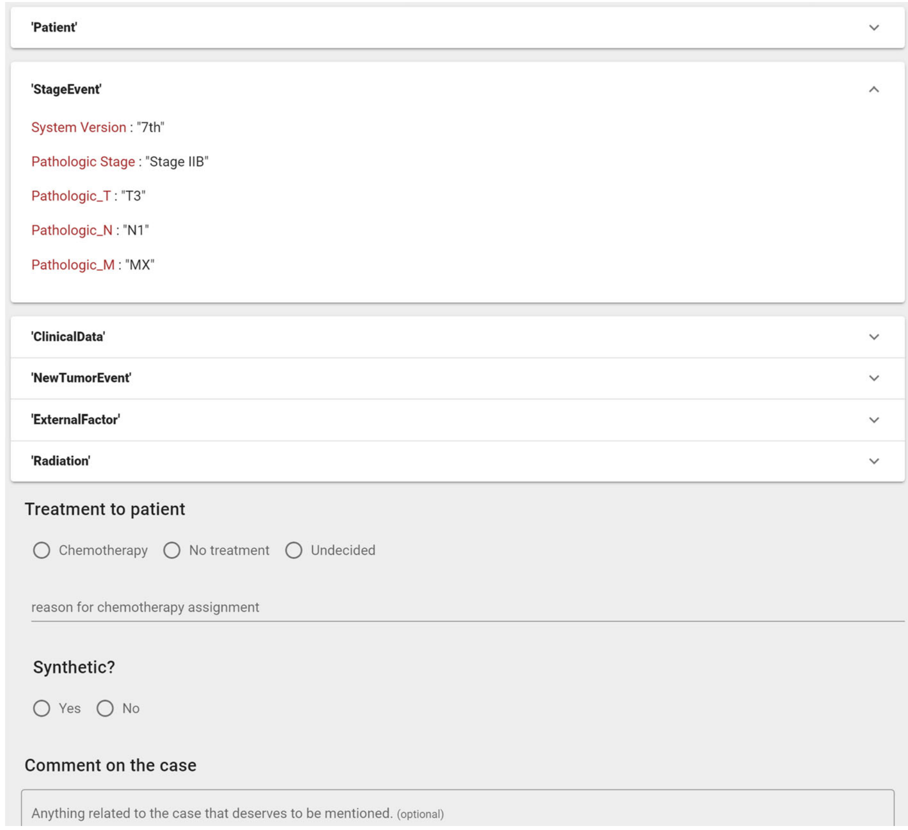


Fig. 5 Web application user interface

pancreatic cancer). As explained in Sect. 4, the cases were selected from among those featuring the highest uncertainty and those generated by the CTGAN (i.e. synthetic). The experts were asked to annotate the new cases regarding whether or not to start chemotherapy and to complete a comments field with any observations, and were also asked whether they considered the patient data to be real or synthetic.

6.2 Context of use analysis

Analysing the context of use is important because it helps to properly design the study and interpret its results. We considered three different aspects (see Fig. 6) that define the context: user, task, and environment. These attributes and their associated subattributes are described in detail in Alonso-Ríos et al. [5].

From the point of view of the user, note that the human experts interacted directly with the system, and even though the participants were not familiar with this kind of system, no technical help was provided in advance. In the design phase, we aimed for a familiar interface (i.e. web application) that fitted perfectly with our purpose of collecting the required data. Each user was an expert in the field, with highly specific domain knowledge, and physical and cognitive characteristics were considered normal. Attitudes to the system were very positive and collaboration by the users was optimal. The domain user expert attributes were not an obstacle, even though the users were not familiar with any similar systems. We consider this not to be an issue as the experts were familiar with web applications and interacted frequently with computers.

The task set for the experts was to read a pancreatic cancer patient case report, with the most relevant attributes of the disease presented by means of a web form. The experts were asked to complete information on the prescribed treatment, making choices based on their expertise, and were asked to determine whether the patient data reflected a real or synthetic case.

The web application created to obtain the annotated data from the experts was the result of several iterations involving a heuristic evaluation of the most important usability aspects derived from the chosen taxonomy [4]. For two of the topics, treatment and real/synthetic case identification, radio buttons were provided in the form, as only one answer was possible. Two text components allowed the expert to include subjective comments on the treatment and the case in general.

Task complexity and frequency were low, with users only completing the task three times in one month. Time taken to complete the task varied greatly, depending on the case. Medical diagnosis is a complex process and,

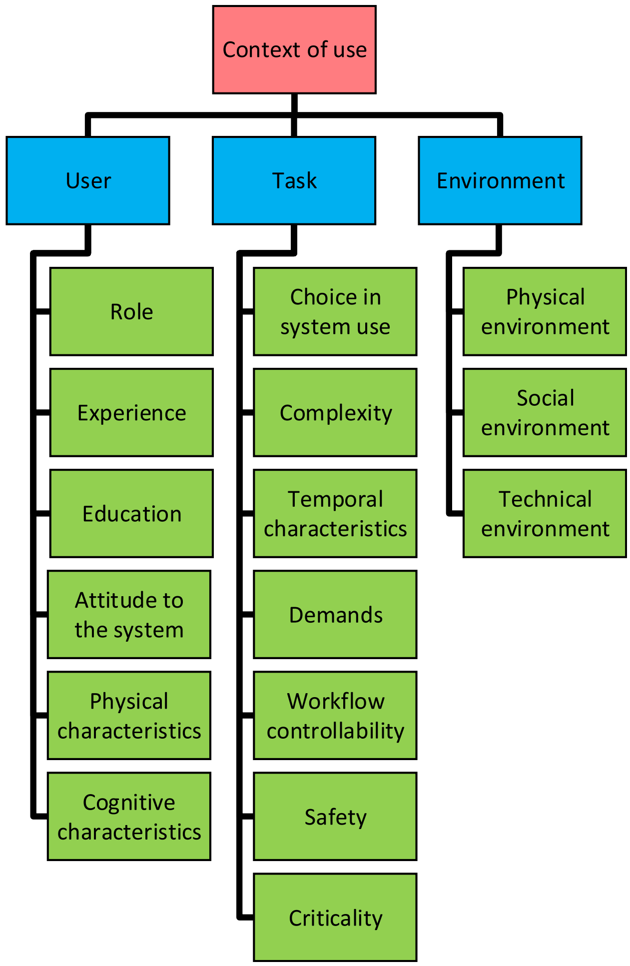


Fig. 6 Context of use main attributes

depending on the evidence and patient data, may require more than five minutes for a single case. Only ten patient cases were therefore presented in each session to ensure sustained motivation.

Finally, the environment consisted of a simple personal computer setup in an office, with no obstacles in terms of sensorial, atmospheric, spatial, or safety conditions. Technological requirements were completely fulfilled and no issues were detected. The *social environment* could affect the focus on the task, and the experts could even be interrupted. Even though they must interact with the system individually, we do not consider human collaboration as an obstacle, as it can only be beneficial for the accuracy on the answers.

6.3 Usability analysis

For our usability study of the application, we applied several approaches in order to obtain the maximum information without interfering with user task completion. The usability study was carefully designed taking into account the characteristics and limitations of our specific context of use, as outlined above.

As described by Ivory et al. [34], several usability evaluation methods and techniques exist that can be used in combination or alone. Our analysis was based on three classes, namely, *inspection*, whereby evaluators apply a set of criteria or heuristics to identify usability problems, *testing*, whereby users perform tasks with the application in the setting described above, and finally, *inquiry*, whereby subjective opinions on testing are collected.

6.3.1 Inspection

As shown by Munro [59], an application for annotating examples—such as that described in this paper—needs to be carefully designed to ensure effectiveness. Inspection usability techniques are crucial in this regard and are typically employed in the system design and initial implementation stages. Of the many types of inspection techniques, one of the most widely used is heuristic evaluation [64], i.e. according to rules of thumb, as it provides quick and easy heuristics for designing an interface. Perhaps the most popular heuristics are the ones proposed by Nielsen [63] (e.g. aesthetic and minimalist design, flexibility, and efficiency of use).

Our goal was to give the users a fully functional application from the outset, rather than an initial prototype to be refined over several cycles. Before actual user testing began, therefore, we performed a heuristic evaluation of the application based on the framework proposed in Alonso-Ríos et al. [6] that proposes an initial systematic and generalizable approach to heuristic evaluation that is then explicitly connected to, and extends, Nielsen's heuristics [63].

After several iterations of finding and fixing usability problems, we obtained an application that could be used in testing with actual users in their routine working environment.

6.3.2 Testing

The experiment consisted of having users (with great expertise in their domain but not necessarily with computers) interacting with a real application requiring them to annotate patient cases. Note that the interface built was a real web application, and no prototypes were discussed or A/B testing was performed. The environment was a real setup where the medical doctors interacted with the application on their own. The aim was to avoid any interference, as in a real use case, so no execution times were recorded (the task needed to be correctly performed, so time was not relevant), and the sessions were not recorded.

Instructions on how to use the web application were not provided as we wanted to test the intuitiveness of the tool in which a simple task needed to be performed. Complexity was intrinsically related to the pancreatic cancer evaluation. We expected *formal use* of the application, which required completing the specific task of selecting suitable treatment, declaring whether the case was real or synthetic, and briefly describing the selected treatment or the case in general.

6.3.3 Inquiry

Inquiry methods require the user to provide feedback on an interface via interviews and/or surveys. In our case we prepared and distributed a questionnaire to the users. We based this on the usability taxonomy proposed by Alonso-Ríos et al. [4], due to its comprehensiveness and clarity, and how it relates with the context of use analysis described above.

Questionnaire literature was also consulted before we produced our own questionnaire, particularly the USE Questionnaire [48], the Software Usability Measurement Inventory (SUMI) [43] and the Cognitive Dimensions framework [11], and we also consulted a review on the applicability of these and other questionnaires by Hinderks et al. [27]. Our goal was to cover the most significant usability aspects related to the studied problem without burdening the users with an excessive number of questions, as this could act as a disincentive to collaboration.

The usability taxonomy has previously been used for usability questionnaires custom-built for a different domain (e.g. [3, 71, 72]). Since the taxonomy is generic and consists of dozens of subattributes, the first step was to remove the attributes that were not applicable to our study. The fact that the taxonomy is hierarchically structured facilitates the pruning of branches of attributes and helps to focus on the relevant usability criteria. Due to the limited availability of the domain user experts, our priority was to ensure that the questionnaire was very brief, so we only included what we considered to be the most essential usability questions.

The generic attributes were progressively refined to obtain more specific subattributes that populated subsequent taxonomic levels (see Fig. 7). Based on this structured taxonomy, we prepared the eleven questions described in Table 2. The taxonomic categories covered by each question are listed contiguously in a separate column.

Each of the eleven questions was answered with a value between 1 and 5, where 1 represented maximum disagreement and 5 maximum satisfaction. A comments field was also provided for users to submit feedback.

Questions 1 and 2 focused on knowability, a property by means of which the user can understand, learn, and remember how to use the system. As the web application was used by the human experts without any prior technical instruction, it was important to capture their thoughts in this regard. Questions 3, 4, and 5 covered operability issues, from the completeness of the tool, to its flexibility in terms of workflow, passing through terminology and cultural aspects (i.e. universality). Questions 6 and 7 referred to efficiency, reflecting task complexity in terms of mental effort over inherent task complexity. Question 8 covered robustness to internal error. Question 9 referred to the safety of the system, particularly in terms of preventing legal issues. Question 10 covered the user's subjective satisfaction and interest. Finally, Question 11 was an overall usability question whose answer should match the answers to the previous questions.

7 Results

7.1 Training results

To measure model performance, we needed to closely examine the evolution of the accuracy value during the different iterations. Since our initial data contained only 181 cases (some initial cases were discarded because they were incomplete), we used a cross-validation strategy to obtain an accuracy value that minimized randomness in the selection of the training and test sets.

Our project baseline was the model trained without following a HITL strategy. Figure 8a shows the result for this baseline model, with overall accuracy of around 60%. Figure 8b shows the result for the HITL strategy, which obtained accuracy of close to 75%, representing a substantial improvement for so few data. To avoid possible overfitting in the final training of the model, an *early stopping* strategy was followed.

The HITL experiment consisted of three iterations, considered the ideal number of iterations both to check if our strategy had an effect on learning, but also not to overburden the physicians. As commented in Sect. 6 and later in the conclusions, a HITL strategy should always take into account HCI issues.

After the three iterations, we ended up with a dataset of 292 cases, 30 of them labelled by the experts [(ten in each iteration), selected by the sampling strategy explained in Sect. 4]. Figure 9 shows how, at each iteration, the classification performance of the model improved even despite the small number of annotated cases added.

We consider that, if more iterations had been performed, accuracy would have been improved further, although would likely have tapered off. Most cases added in each iteration were synthetic cases, variations of the cases included in the initial limited dataset. Note that more data

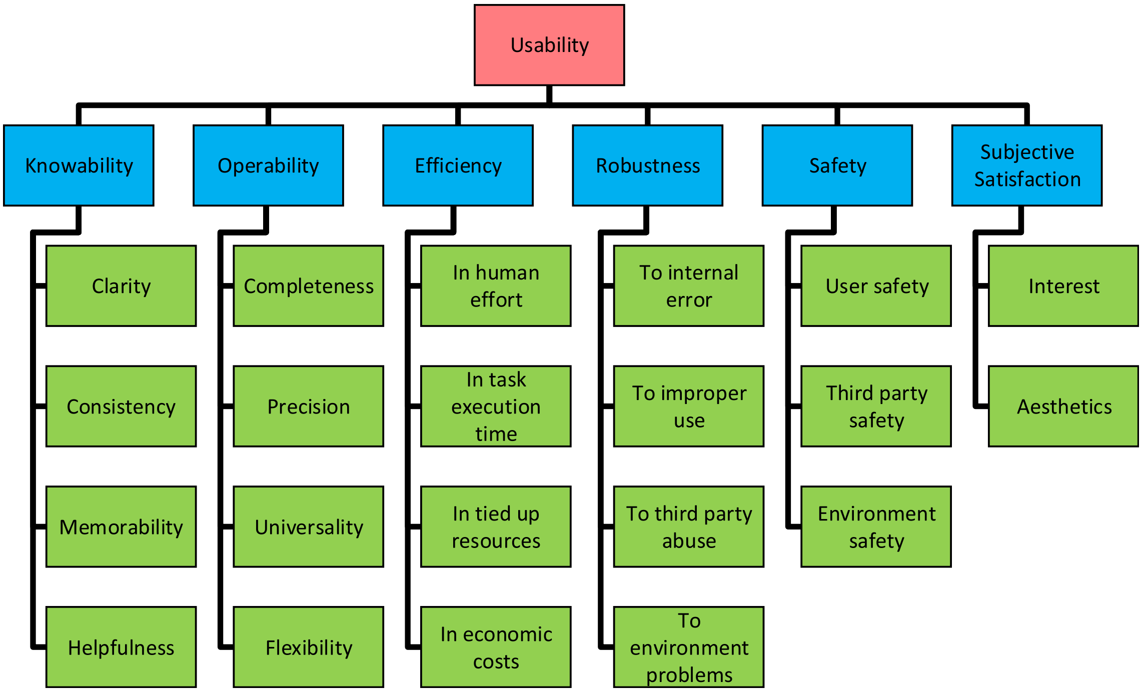


Fig. 7 Usability main attributes

Table 2 Usability questionnaire

Question	Taxonomy category
(1) I understand the process to input data in the system	(K): Clarity in functioning/User tasks
(2) I understand the answers provided by the system	(K): Clarity in functioning/System tasks
(3) The system provides everything I need to be able to use it	(O): Completeness
(4) The terminology used by the system seems correct to me	(O): Universality/Cultural universality
(5) The flexibility in using the system seems correct to me	(O): Flexibility/controllability/Workflow controllability
(6) I do not need to invest special mental efforts to use the system	(E): In human effort/Mental
(7) I do not need to spend too much time using the system	(E): In task execution time
(8) The system looks robust and I do not detect potential issues	(R): Robustness to internal error
(9) The system complies with current regulations	(S): User safety/Legal safeguarding
(10) I consider the system useful and interesting	(SS): Interest
(11) Overall I consider the system easy to use	Usability

*Categories: Knowability (K), Operability (O), Efficiency (E), Robustness (R), Safety (S), Subjective satisfaction (SS)

implies a greater workload for the physicians and a greater computational workload.

7.2 Usability questionnaire results

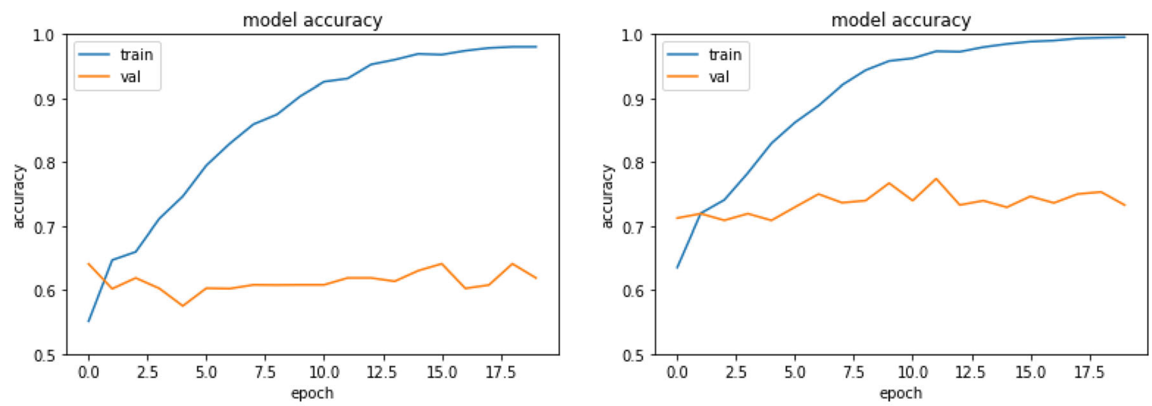
As part of the usability study, we analysed the results of the questionnaire distributed to the domain experts (see Table 2).

Maximum scores awarded were 4 for questions 1 and 2, and 5 for the remaining nine questions. Therefore, while the web application could be considered successful in terms of usability, there were two questions that received less than the highest score, both related to the knowability attribute, and in particular, with clarity in functioning from both the user and system perspectives. Those were:

Q1. *I understand the process to input the data in the system.*

Q2. *I understand the answers provided by the system.*

Knowability is defined as the user understanding, learning and remembering how to use the system. With many ML models, it is a real challenge to provide the means by which



(a) Baseline model result (b) HITL model result

Fig. 8 Comparison of results before and after deploying the HITL strategy

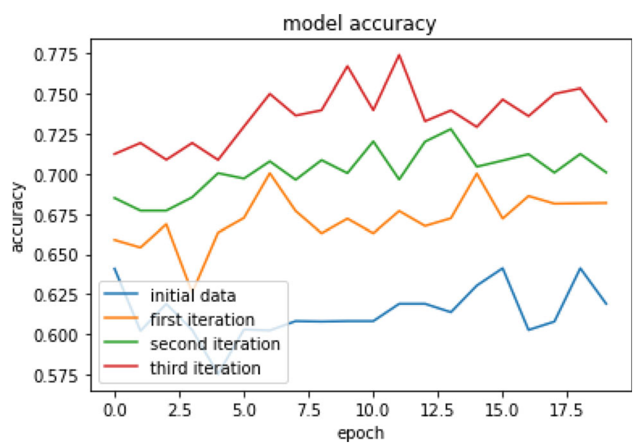


Fig. 9 Three iterations in HITL training compared

a final user can understand the internal workings of the models.

Since the medical experts did not complete the free-form comments section of the questionnaire for those two questions, we requested more details in order to understand why the scores were below the maximum. The general comment was that it was not clear how the system used the answers provided by them, nor how the cases were chosen or annotated before human interaction with the system. We also relate their uncertainty to the fact that no precise instructions were issued. Nevertheless, the system was simple enough for the users to be able to complete the task.

As future work, a different type of model (e.g. decision tree) could be used so that the final users could better understand the more important features and how they are related to each other. An explainability method could also be applied to the current model so that, in order that they understand the underlying logic, the most relevant attributes are presented to the final user.

8 Discussion and conclusions

Data bottlenecks are a problem within ML, especially in complicated domains such as medical environments where there are either few data, or what data exist are not labelled, or data have weak or unreliable labels.

One solution to alleviate data bottlenecks is to use data augmentation techniques, which generate synthetic cases that make it easier for the ML model to learn existing patterns in the data. While these techniques are easy to apply when dealing with images, but less so when working with tabular data, as the data contain relationships between the values of the different features that the synthetic case generator needs to take into account. CTGANs are generally used to generate synthetic data from tabular data, but in a data-poor environment, they share the problem of scarce data, meaning that they cannot learn from the relationships between features and take them into account when generating synthetic cases.

This is where HITL techniques can address the data bottleneck problem. In an AL process, weak labels—such as those generated for synthetic cases, and even labels from the dataset itself—are analysed and corrected by human experts. Humans, even if they analyse just a few cases, can have a significant impact on system accuracy, as recently demonstrated in Gupta and Sintorn [24], Bravo-Rocca et al. [13], Zhao et al. [92], and Khanal et al. [42].

Humans, however, can go beyond merely labelling cases. In our particular case, humans also analysed cases to decide whether or they were synthetic, justified their decision and, if the case was real, provided a rationale as to why it was assigned a particular label.

Reasons why human experts considered a case to be synthetic were used to create new constraints for the CTGAN model. Thus, humans acted as an additional discriminatory layer, thereby enabling the CTGAN to generate better synthetic cases that were more indistinguishable from real cases and, therefore, more suitable for the learning process.

Basically what the human experts did was solving one of the problems of the data augmentation process, which is to assess and evaluate the quality of augmented datasets. As deployment of data augmentation methods grows, so also will the need to analyse output quality. Furthermore, human experts can help detect and correct whatever biases may have been carried over from the original dataset to data supplemented from that dataset.

However, including humans in the learning process has a cost, especially when dealing with domain experts whose availability may be restricted, as happens in the medical domain. Since having HCI experts is essential to ensure the success of experiments like ours, this makes it imperative to take HCI-related aspects into account when designing any experiment that includes them.

The idea is to reach a trade-off between the amount of cases an expert can collaborate on (the more the better), and the amount of time and effort the expert will invest in analysing those cases (the less the better). A good design of the user interface and the user interaction is especially important to maintaining the user's interest and collaboration and avoiding excessive demands on them [54].

An important event demonstrating how human behaviour can influence HITL experiments occurred with the explanation of why a case was labelled in a given category (the second component in the expert responses). Our intention was to use this information to improve the explanatory capacities of the system, yet the expert responses were less detailed than responses given when identifying synthetic cases, which ultimately means that the responses were not very useful. We believe that the reason is that the experts invested most effort in detecting whether a case was synthetic or not because they were keen not to be "fooled" by the machine (just as we carefully watch a magician to try and discover the trick and so fail to pay attention to the rest of the show). A possible solution would be to design the iterations differently, e.g. include an iteration in which the experts are aware that all the cases are synthetic, so they simply provide explanations as to why the data are synthetic, and another iteration in which the experts are aware that all the cases are real, so they focus only on explaining their labels.

8.1 Future work

Several options are being considered for future work. Firstly, regarding the IML process that includes new restrictions in the CTGAN, the process is currently manual, because expert opinions are collected in natural language that is ambiguous and often needs additional clarifications. A possible improvement would be to incorporate either a natural language processor that can generate constraints automatically, or a more complex interface that allows physicians to set system constraints interactively. Either of these solutions would be quite complex; in the first case, there would be no guarantee that the constraint built from natural language was exactly what the expert meant, and in the second case, the interactive tool would add complexity, would require a more extensive study of usability, and would not guarantee that all the possible constraints expressed by experts could be represented, not to mention the additional demands on the experts.

Secondly, since IML has been successfully applied in applications dealing with unstructured data (using experts to give structure to such data), it could also be applied to interactive image segmentation. The aim is to simplify the process of eliciting knowledge by involving experts as users of the IML tool in annotating image content that is relevant to the model. According to Holzinger et al. [31], IML is especially suited for applications in the medical field.

Finally, by incorporating human knowledge and skills we not only improve the quality of the learning models and build them with fewer data, but can also use this knowledge to enable features such as retraceability and explainability that could mitigate the black-box problem in certain ML models. Also useful is the possibility for comparing the explanations of human experts with the explanations obtained by transparent ML models such as decision trees.

Acknowledgements This work has been supported by the State Research Agency of the Spanish Government (Grant PID2019-107194GB-I00/AEI/10.13039/501100011033) and by the Xunta de Galicia (Grant ED431C 2022/44), supported in turn by the EU European Regional Development Fund. We wish to acknowledge support received from the Centro de Investigación de Galicia CITIC, funded by the Xunta de Galicia and the European Regional Development Fund (Galicia 2014–2020 Program; Grant ED431G 2019/01).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Availability of data and materials The dataset analysed in this study is available in the TCGA repository [83], https://portal.gdc.cancer.gov/projects/TCGA-PAAD.

Declarations

Conflict of interest The authors declare that they have no conflict of interest regarding the publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abdar M, Mehrzadi A, Goudarzi M et al (2023) Binarized multi-gate mixture of Bayesian experts for cardiac syndrome x diagnosis: a clinician-in-the-loop scenario with a belief-uncertainty fusion paradigm. Inf Fusion 97(101):813. https://doi.org/10.1016/j.inffus.2023.101813
2. Adelman L, Riedel SL (2012) Handbook for evaluating knowledge-based systems: conceptual framework and compendium of methods. Springer, Berlin
3. Alonso-Ríos D, Raneburger D, Popp R et al (2014) A user study on tailoring GUIs for smartphones. In: Proceedings of the 29th annual ACM symposium on applied computing. Association for Computing Machinery, New York, NY, USA, SAC '14, pp 186–192. https://doi.org/10.1145/2554850.2555085
4. Alonso-Ríos D, Vázquez-García A, Mosqueira-Rey E et al (2009) Usability: a critical analysis and a taxonomy. Int J Hum Comput Interact 26(1):53–74. https://doi.org/10.1080/10447310903025552
5. Alonso-Ríos D, Vázquez-García A, Mosqueira-Rey E et al (2010) A context-of-use taxonomy for usability studies. Int J Hum Comput Interact 26(10):941–970. https://doi.org/10.1080/10447318.2010.502099
6. Alonso-Ríos D, Mosqueira-Rey E, Moret-Bonillo V (2018) A systematic and generalizable approach to the heuristic evaluation of user interfaces. Int J Hum Comput Interact 34(12):1169–1182. https://doi.org/10.1080/10447318.2018.1424101
7. Amershi S, Cakmak M, Knox WB et al (2014) Power to the people: the role of humans in interactive machine learning. AI Mag 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513
8. Aroyo L, Lease M, Paritosh P, Schaekermann M (2022) Data excellence for AI: Why should you care? Interactions 29(2):66–69. https://doi.org/10.1145/3517337
9. Belete DM, Huchaiah MD (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl 44(9):875–886. https://doi.org/10.1080/1206212X.2021.1974663
10. Bengio Y, Louradour J, Collobert R et al (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. Association for Computing Machinery, New York, NY, USA, ICML '09, pp 41–48. https://doi.org/10.1145/1553374.1553380
11. Blackwell AF, Britton C, Cox AL et al (2001) Cognitive dimensions of notations: design tools for cognitive technology. In: Proceedings of the 4th international conference on cognitive technology: instruments of mind. Springer-Verlag, Berlin, Heidelberg, CT '01, pp 325–341
12. Boecking B, Neiswanger W, Xing E et al (2021) Interactive weak supervision: Learning useful heuristics for data labeling. arXiv:2012.06046
13. Bravo-Rocca G, Liu P, Guitart J et al (2022) Human-in-the-loop online multi-agent approach to increase trustworthiness in ml models through trust scores and data augmentation. In: 2022 IEEE 46th annual computers, software, and applications conference (COMPSAC), pp 32–37. https://doi.org/10.1109/COMPSAC54236.2022.00014
14. Chen L, Wang J, Guo B et al (2023) Human-in-the-loop machine learning with applications for population health. CCF Trans Pervasive Comput Interact 5(1):1–12. https://doi.org/10.1007/s42486-022-00115-4
15. Choi E, Biswal S, Malin B et al (2017) Generating multi-label discrete patient records using generative adversarial networks. In: Doshi-Velez F, Fackler J, Kale D et al (eds) Proceedings of the 2nd machine learning for healthcare conference, proceedings of machine learning research, vol 68. PMLR, pp 286–305. https://proceedings.mlr.press/v68/choi17a.html
16. Delussu R, Putzu L, Fumera G (2023) Human-in-the-loop cross-domain person re-identification. Expert Syst Appl 226(120):216. https://doi.org/10.1016/j.eswa.2023.120216
17. Donmez P, Carbonell JG (2008) Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In: Proceedings of the 17th ACM conference on information and knowledge management. Association for Computing Machinery, New York, NY, USA, CIKM '08, pp 619–628. https://doi.org/10.1145/1458082.1458165,
18. Fiebrink R, Cook PR (2010) The wekinator: a system for real-time, interactive machine learning in music. In: Proceedings of the eleventh international society for music information retrieval conference (ISMIR 2010), Utrecht
19. Fu J, Li W, Du J et al (2021) Dsagan: a generative adversarial network based on dual-stream attention mechanism for anatomical and functional image fusion. Inf Sci 576:484–506. https://doi.org/10.1016/j.ins.2021.06.083
20. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C et al (eds) Advances in neural information processing systems, vol 27. Curran Associates Inc., New York
21. Gray WD, Salzman MC (1998) Damaged merchandise? A review of experiments that compare usability evaluation methods. Hum Comput Interact 13(3):203–261
22. Gunning D (2017) Explainable artificial intelligence (xAI). Tech. rep., Defense Advanced Research Projects Agency (DARPA). https://www.darpa.mil/program/explainable-artificial-intelligence
23. Guo K, Hu X, Li X (2022) Mmfgan: a novel multimodal brain medical image fusion based on the improvement of generative adversarial network. Multimedia Tools Appl 81(4):5889–5927
24. Gupta A, Sintorn IM (2023) Towards better guided attention and human knowledge insertion in deep convolutional neural networks. In: Karlinsky L, Michaeli T, Nishino K (eds) Computer vision–ECCV 2022 workshops. Springer, Cham, pp 437–453
25. Halder A, Kumar A (2019) Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data. J Biomed Inform 92(103):136. https://doi.org/10.1016/j.jbi.2019.103136
26. Hills TT, Todd PM, Lazer D et al (2015) Exploration versus exploitation in space, mind, and society. Trends Cogn Sci 19(1):46–54. https://doi.org/10.1016/j.tics.2014.10.004
27. Hinderks A, Winter D, Schrepp M et al (2019) Applicability of user experience and usability questionnaires. J Univ Comput Sci 25(13):1717–1735
28. Holmberg L, Davidsson P, Linde P (2020) A feature space focus in machine teaching. In: 2020 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), pp 1–2. https://doi.org/10.1109/PerComWorkshops48775.2020.9156175. http://mau.diva-portal.org/smash/get/diva2:1428195/FULLTEXT01.pdf
29. Holzinger A (2016) Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform 3(2):119–131. https://doi.org/10.1007/s40708-016-0042-6
30. Holzinger A, Biemann C, Pattichis CS et al (2017) What do we need to build explainable AI systems for the medical domain? arXiv e-prints arXiv:1712.09923 [cs.AI]

31. Holzinger A, Plass M, Kickmeier-Rust M et al (2019) Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl Intell 49(7):2401–2414. https://doi.org/10.1007/s10489-018-1361-5
32. Hong DS, Baik C (2021) Generating and validating synthetic training data for predicting bankruptcy of individual businesses. J Inf Commun Converg Eng 19(4):228–233
33. ISO 9241-210:2010 (2010) ISO 9241-210:2010—ergonomics of human–system interaction—Part 210: Human-centred design for interactive systems
34. Ivory MY, Hearst MA (2001) The state of the art in automating usability evaluation of user interfaces. ACM Comput Surv 33(4):470–516. https://doi.org/10.1145/503112.503114
35. Jeong JJ, Tariq A, Adejumo T et al (2022) Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. J Digit Imaging 35:1–16
36. Jia J, Wu P, Zhang K et al (2022) Imbalanced disk failure data processing method based on CTGAN. In: Intelligent computing theories and application: 18th international conference, ICIC 2022, Xi'an, China, August 7–11, 2022, Proceedings, Part II. Springer, Berlin, Heidelberg, pp 638–649. https://doi.org/10.1007/978-3-031-13829-4_55
37. Jiang J, Hu YC, Tyagi N et al (2018) Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 777–785
38. Jiang M, Zhi M, Wei L et al (2021) FA-GAN: fused attentive generative adversarial networks for MRI image super-resolution. Comput Med Imaging Graph. https://doi.org/10.1016/j.compmedimag.2021.101969
39. Ju J, Wismans LV, Mustafa DA et al (2021) Robust deep learning model for prognostic stratification of pancreatic ductal adenocarcinoma patients. Iscience 24(12):1–18
40. Kabra M, Robie AA, Rivera-Alba M et al (2013) Jaaba: interactive machine learning for automatic annotation of animal behavior. Nat Methods 10(1):64–67. https://doi.org/10.1038/nmeth.2281
41. Khan W, Zaki N, Ahmad A et al (2022) Mixed data imputation using generative adversarial networks. IEEE Access 10:124,475-124,490. https://doi.org/10.1109/ACCESS.2022.3218067
42. Khanal S, Refati R, Glandt K et al (2021) Using content analysis and machine learning to identify Covid-19 information relevant to low-income households on social media. In: 2021 IEEE international conference on parallel and distributed processing with applications, big data and cloud computing, sustainable computing and communications, social computing and networking (ISPA/BDCloud/SocialCom/SustainCom), pp 1522–1531. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00205
43. Kirakowski J, Corbett M (1993) SUMI: the software usability measurement inventory. Br J Educ Technol 24(3):210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x
44. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
45. Lee S, Amgad M, Masoud M et al (2019) An ensemble-based active learning for breast cancer classification. In: 2019 IEEE international conference on bioinformatics and Biomedicine (BIBM), pp 2549–2553. https://doi.org/10.1109/BIBM47256.2019.8983317
46. Lewis JR (2014) Usability: lessons learned... and yet to be learned. Int J Hum Comput Interact 30(9):663–684
47. Lison P, Hubin A, Barnes J et al (2020) Named entity recognition without labelled data: a weak supervision approach. arXiv:2004.14723
48. Lund AM (2001) Measuring usability with the use questionnaire. Usab Interface 8(2):3–6
49. Mendes J, Pereira T, Silva F et al (2023) Lung CT image synthesis using GANS. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.119350
50. Meza Martínez MA, Nadj M, Maedche A (2019) Towards an integrative theoretical framework of interactive machine learning systems. In: Proceedings of the 27th European conference on information systems (ECIS), Stockholm, Uppsala, Sweden. https://aisel.aisnet.org/ecis2019_rp/172
51. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
52. Moon J, Jung S, Park S et al (2020) Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting. IEEE Access 8:205,327-205,339. https://doi.org/10.1109/ACCESS.2020.3037063
53. Moreno Planas JM, Sánchez Ortega A, García Bueno JM et al (2010) Avances en Cáncer de Páncreas, del laboratorio a la clínica. UCLM, AECC
54. Mosqueira-Rey E, Alonso-Ríos D, Baamonde-Lozano A (2021) Integrating iterative machine teaching and active learning into the machine learning loop. Procedia Comput Sci 192:553–562. https://doi.org/10.1016/j.procs.2021.08.057
55. Mosqueira-Rey E, Fernández-Castaño S, Alonso-Ríos D et al (2023a) Gamifying machine teaching: human-in-the-loop approach for diphthong and hiatus identification in Spanish language. In: knowledge-based and intelligent information and engineering systems: proceedings of the 27th international conference KES2023
56. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D et al (2023) Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 56:3005–3054. https://doi.org/10.1007/s10462-022-10246-w
57. Mottini A, Lhéritier A, Acuna-Agost R (2018) Airline passenger name record generation using generative adversarial networks. arXiv:1807.06657
58. Mullainathan S, Obermeyer Z (2022) Solving medicine's data bottleneck: nightingale open science. Nat Med 28(5):897–899. https://doi.org/10.1038/s41591-022-01804-4
59. Munro R (2020) Human-in-the-loop machine learning. Manning Publications, New York
60. Na J, Kim SJ, Kim H et al (2023) A unified microstructure segmentation approach via human-in-the-loop machine learning. Acta Mater 255(119):086. https://doi.org/10.1016/j.actamat.2023.119086
61. NCCN (2019) Pancreatic adenocarcinoma, version 3.2019. National Comprehensive Cancer Network. http://pancreatic.altervista.org/downloads/NCCN3.2019Pancreatic.pdf
62. Nicolle R, Raffenne J, Paradis V et al (2019) Prognostic biomarkers in pancreatic cancer: avoiding errata when using the TCGA dataset. Cancers 11(1):126
63. Nielsen J (2020) 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/. Accessed: 01 May 2023
64. Nielsen J, Molich R (1990) Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp 249–256
65. Nugraha RA, Parede HF, Subekti A (2022) Oversampling based on generative adversarial networks to overcome imbalance data in predicting fraud insurance claim. Kuwait J Sci. https://doi.org/10.48129/kjs.splml.19119
66. Park N, Mohammadi M, Gorde K et al (2018) Data synthesis based on generative adversarial networks. Proc VLDB Endow 11(10):1071–1083. https://doi.org/10.14778/3231751.3231757
67. Patki N, Wedge R, Veeramachaneni K (2016) The synthetic data vault. In: 2016 IEEE international conference on data science and advanced analytics (DSAA), pp 399–410. https://doi.org/10.1109/DSAA.2016.49

68. Porter R, Theiler J, Hush D (2013) Interactive machine learning in data exploitation. Comput Sci Eng 15(5):12–20. https://doi.org/10.1109/MCSE.2013.74
69. Rahman AU, Al-Obeidat F, Tubaishat A et al (2022) Discovering the correlation between phishing susceptibility causing data biases and big five personality traits using C-GAN. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2022.3201153
70. Ramos G, Meek C, Simard P et al (2020) Interactive machine teaching: a human-centered approach to building machine-learned models. Hum Comput Interact 35(5–6):413–451. https://doi.org/10.1080/07370024.2020.1734931
71. Raneburger D, Alonso-Ríos D, Popp R et al (2013) A user study with GUIs tailored for smartphones. In: Kotzé P, Marsden G, Lindgaard G et al (eds) Human–computer interaction—INTERACT 2013. Springer, Berlin, Heidelberg, pp 505–512
72. Raneburger D, Popp R, Alonso-Ríos D et al (2013b) A user study with GUIs tailored for smartphones and tablet PCs. In: 2013 IEEE international conference on systems, man, and cybernetics, pp 3727–3732. https://doi.org/10.1109/SMC.2013.635
73. Rubens N, Elahi M, Sugiyama M et al (2015) Active learning in recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer, Berlin, pp 809–846. https://doi.org/10.1007/978-1-4899-7637-6_24
74. Ryan M (2020) Deep learning with structured data. Manning Publications Co, New York
75. Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2023) Application of streaming analytics for artificial lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps. Neural Comput Appl 35(2):1247–1277. https://doi.org/10.1007/s00521-022-07995-8
76. Settles B (2009) Active learning literature survey. Technical report, University of Wisconsin-Madison. Department of Computer Sciences. https://minds.wisconsin.edu/handle/1793/60660
77. Shani C, Zarecki J, Shahaf D (2023) The lean data scientist: recent advances toward overcoming the data bottleneck. Commun ACM 66(2):92–102. https://doi.org/10.1145/3551635
78. Shin HC, Tenenholtz NA, Rogers JK et al (2018) Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International workshop on simulation and synthesis in medical imaging. Springer, pp 1–11
79. Simard PY, Amershi S, Chickering DM et al (2017) Machine teaching: a new paradigm for building machine learning systems. arXiv e-prints arXiv:1707.06742
80. Sinkala M, Mulder N, Martin D (2020) Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. Sci Rep 10(1):1212
81. Tang J, Fan B, Xiao L et al (2021) A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. SPE J 26(01):482–497. https://doi.org/10.2118/204224-PA
82. Tlachac M, Gerych W, Agrawal K et al (2022) Text generation to aid depression detection: a comparative study of conditional sequence generative adversarial networks. In: 2022 IEEE international conference on big data (big data), pp 2804–2813. https://doi.org/10.1109/BigData55660.2022.10020224,
83. Tomczak K, Czerwińska P, Wiznerowicz M (2015) The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 19(1A):68–77. https://doi.org/10.5114/wo.2014.47136
84. Wali A, Alamgir Z, Karim S et al (2022) Generative adversarial networks for speech processing: a review. Comput Speech Lang 72(101):308
85. Wang H, Miller DJ, Kesidis G (2023) Anomaly detection of adversarial examples using class-conditional generative adversarial networks. Comput Secur 124:10. https://doi.org/10.1016/j.cose.2022.102956
86. Wang J, Yan X, Liu L et al (2022) Cttgan: traffic data synthesizing scheme based on conditional GAN. Sensors 22(14):10. https://doi.org/10.3390/s22145243
87. Wang Z, She Q, Ward TE (2021) Generative adversarial networks in computer vision: a survey and taxonomy. ACM Comput Surv (CSUR) 54(2):1–38
88. Wen S, Kurc TM, Hou L et al (2018) Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images. AMIA Summits Transl Sci Proc 2018:227
89. Xu L, Skoularidou M, Cuesta-Infante A et al (2019) Modeling tabular data using conditional GAN. In: Wallach H, Larochelle H, Beygelzimer A et al (eds) Advances in neural information processing systems, vol 32. Curran Associates Inc, London
90. Xu W (2019) Toward human-centered AI: a perspective from human–computer interaction. Interactions 26(4):42–46. https://doi.org/10.1145/3328485
91. Zhan B, Li D, Wu X et al (2022) Multi-modal MRI image synthesis via GAN with multi-scale gate mergence. IEEE J Biomed Health Inform 26(1):17–26. https://doi.org/10.1109/JBHI.2021.3088866
92. Zhao Z, Xu P, Scheidegger C et al (2022) Human-in-the-loop extraction of interpretable concepts in deep learning models. IEEE Trans Vis Comput Graph 28(1):780–790. https://doi.org/10.1109/TVCG.2021.3114837
93. Zhu JY, Park T, Isola P et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
94. Zhuang H, Zhang J, Liao F (2021) A systematic review on application of deep learning in digestive system image processing. Vis Comput 39:46–51. https://doi.org/10.1007/s00371-021-02322-z

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Eduardo Mosqueira-Rey[1] · Elena Hernández-Pereira[1] · José Bobes-Bascarán[1] · David Alonso-Ríos[1] · Alberto Pérez-Sánchez[1] · Ángel Fernández-Leal[1] · Vicente Moret-Bonillo[1] · Yolanda Vidal-Ínsua[2] · Francisca Vázquez-Rivera[2]

✉ Eduardo Mosqueira-Rey
eduardo@udc.es

Elena Hernández-Pereira
elena.hernandez@udc.es

José Bobes-Bascarán
jose.bobes@udc.es

David Alonso-Ríos
david.alonso@udc.es

Alberto Pérez-Sánchez
alberto.perez.sanchez@udc.es

Ángel Fernández-Leal
angel.fleal@udc.es

Vicente Moret-Bonillo
vicente.moret@udc.es

Yolanda Vidal-Ínsua
yvidalinsua@gmail.com

Francisca Vázquez-Rivera
francisca.vazquez.rivera@sergas.es

[1] Department of Computer Science and Information Technologies, Universidade da Coruña (CITIC), Campus de Elviña, 15071 A Coruña, Spain

[2] Servicio de Oncología Médica, Complejo Hospitalario (CHUS), Rúa da Choupana, s/n, 15706 Santiago de Compostela, Spain