INTERACTIVE MACHINE LEARNING: A STATE OF THE ART REVIEW

A PREPRINT

Natnael A. WONDIMU*
Department of Artificial Intelligence
Addis Ababa University
Addis Ababa, Ethiopia
natnael.argaw@aau.edu.et

Ubbo VISSER †
Department of Computer Science
University of Miami
Florida, USA
visser@cs.miami.edu

Cédric BUCHE ‡
IRL CROSSING, ENIB
Brittany, France
cedric.buche@enib.fr

July 14, 2022

ABSTRACT

Machine learning has proved useful in many software disciplines, including computer vision, speech and audio processing, natural language processing, robotics and some other fields. However, its applicability has been significantly hampered due its black-box nature and significant resource consumption. Performance is achieved at the expense of enormous computational resource and usually compromising the robustness and trustworthiness of the model. Recent researches have been identifying a lack of interactivity as the prime source of these machine learning problems. Consequently, interactive machine learning (iML) has acquired increased attention of researchers on account of its human-in-the-loop modality and relatively efficient resource utilization. Thereby, a state-of-the-art review of interactive machine learning plays a vital role in easing the effort toward building human-centred models. In this paper, we provide a comprehensive analysis of the state-of-the-art of iML. We analyze salient research works using merit-oriented and application/task oriented mixed taxonomy. We use a bottom-up clustering approach to generate a taxonomy of iML research works. Research works on adversarial black-box attacks and corresponding iML based defense system, exploratory machine learning, resource constrained learning, and iML performance evaluation are analyzed under their corresponding theme in our merit-oriented taxonomy. We have further classified these research works into technical and sectoral categories. Finally, research opportunities that we believe are inspiring for future work in iML are discussed thoroughly.

Keywords interactive machine learning, exploratory machine learning, low resource learning, iML evaluation

1 Introduction

Interactive machine learning (iML) is an active machine learning technique in which models are designed and implemented with human-in-the-loop manner. End-users participate in model building process by iteratively feeding training parameters, inspecting model outputs and providing feedback on intermediate results Arendt et al. [2017], Amershi et al. [2014], Fails and Olsen Jr [2003], Jiang et al. [2019], Holzinger et al. [2018], Boukhelifa et al. [2018], Liu et al. [2017a], Françoise [2020].

The conception of iML dates back to the emergence of query learning where queries are used to learn unknown concepts Ghai et al. [2021], Angluin [1988], Holzinger [2016], Chung et al. [2020], Fogarty et al. [2008]. A model building component in iML framework interacts with Oracles by issuing queries for additional training data or feedback against its intermediate results. iML based methods mainly aspires to build robust Huang et al. [2011], Papernot et al. [2017], Carlini and Wagner [2017], Szegedy et al. [2013a], Sun et al. [2018], Brendel et al. [2017], Guo et al. [2019],

---

Cheng et al. [2019a], Guo et al. [2018], Porter et al. [2013], Ma et al. [2019], Das et al. [2020], Slack et al. [2020], Emamjomeh-Zadeh and Kempe [2017] models by trading-off accuracy for trust Teso and Kersting [2019], Gutzwiller and Reeder [2017], Krause et al. [2016], Ribeiro et al. [2016], Mozina [2018], Gutzwiller and Reeder [2017], Holzinger et al. [2018], Turchetta et al. [2019], Berkenkamp et al. [2016], Sui et al. [2018], Van Den Elzen and Van Wijk [2011], Liu et al. [2017b], Zhao et al. [2018], Mühlbacher et al. [2014] and low resource learning Ambati [2011], Frazier and Riedl [2019], Porter et al. [2013], Preuveneers et al. [2020], Holzinger et al. [2017], Fails and Olsen Jr [2003], Amershi et al. [2011], Tegen et al. [2020], Amershi et al. [2012], Dzyuba et al. [2014], Li et al. [2018], Jain et al. [2020], Fiebrink and Cook [2010], Gillian and Paradiso [2014], Schedel et al. [2011], Diaz et al. [2019], Arendt et al. [2018, 2017].

iML is becoming the center of machine learning research Jiang et al. [2019], Arendt et al. [2017] as the need to operate in resource constrained environments and engaging the human-in-the-loop increased over time. Various iML systems have been designed and implemented to alleviate the long-standing setbacks of standard machine learning algorithms. Its impact has been studied in health Holzinger et al. [2018], Wallace et al. [2012], Holzinger [2016], Maadi et al. [2021], Holzinger et al. [2017], Kose et al. [2015], Qian et al. [2011], Tyagi et al. [2018], agriculture de Medeiros et al. [2020], Flutura et al. [2020], Liakos et al. [2018], finance Kose et al. [2015], White [1993], politics and sociology Amershi et al. [2012], Gillies [2019], military and robotics Gutzwiller and Reeder [2017], Berkenkamp et al. [2016], Kyrarini et al. [2019], cyber security Chung et al. [2020], Huang et al. [2011], Papernot et al. [2017], Carlini and Wagner [2017], Brendel et al. [2017], Guo et al. [2019], Slack et al. [2020], Emamjomeh-Zadeh and Kempe [2017], Ilyas et al. [2018], Cheng et al. [2019b], education Witten and Frank [2002], Hollan et al. [1984], Zhou et al. [2020] and game and entertainment Frazier and Riedl [2019], Diaz et al. [2019], Fiebrink and Cook [2010], Schedel et al. [2011].

Moreover, iML has been analysed and built over the predominant standard machine learning algorithms such as SVM Teso and Kersting [2019], Talbot et al. [2009], Tong and Chang [2001], Tsai et al. [2009], Javidi et al. [2008], Tyagi et al. [2018], Talbot et al. [2009], Genetic Algorithms Gutzwiller and Reeder [2017], Kim and Cho [2000], Lai and Chen [2011], Cho and Lee [2002], Wang et al. [2005], Poirson et al. [2020], Ant Colony Holzinger et al. [2019], Meng et al. [2021], Yang et al. [2020] and some other algorithms Tyagi et al. [2018], Frazier and Riedl [2019], Fogarty et al. [2008], Breve [2020], Ha and Eck [2017]. Similarly, its capacity to uncover the details behind black-box models is presented in various research works Gutzwiller and Reeder [2017], Teso and Kersting [2019], Van Den Elzen and Van Wijk [2011], Liu et al. [2017b], Ma et al. [2019], Mühlbacher et al. [2014]. Most research focus ranges from extending standard machine learning models with interactive capabilities to formulating robust performance evaluation mechanisms.

Sate-of-the-art review gives insights on recent progresses on a specific subject matter and paves ways for more research. To this end, numerous surveys have been conducted on the application of iML to specific sectors. Architectural Dudley and Kristensson [2018], Chatzimparmpas et al. [2020], Meza Martínez et al. [2019] and application/sector oriented Jiang et al. [2019], Liakos et al. [2018], Maadi et al. [2021] analysis has been significantly used techniques in the past. According to the meta-analysis Meza Martínez et al. [2019], Dudley and Kristensson [2018], Maadi et al. [2021], Liakos et al. [2018], Chatzimparmpas et al. [2020], Jiang et al. [2019], Amershi et al. [2014], Liu et al. [2017a], Teso and Hinz [2020] we conducted, most surveys lack extensiveness and inclusiveness, leaving the various researches that constitute the state-of-the-art of iML untouched. In Amershi et al. [2014], Dudley and Kristensson [2018] and Meza Martínez et al. [2019], the advancements made towards designing usable interfaces for iML is discussed. A review of iML and interactive visualization has been conducted in Liu et al. [2017a] and Meza Martínez et al. [2019]. Moreover, a review of iML inspired researches in medical and agricultural sectors is extensively studied in Maadi et al. [2021] and Liakos et al. [2018] respectively. A recent task-oriented extensive survey on interactive machine learning is also presented in Jiang et al. [2019].

However, to the best of our knowledge, there has been a little extensive survey in the field of iML covering the state-of-the-art. Besides, no survey offer a merit-oriented and detailed state-of-the-art review. Therefore, the primary motivation of this paper is to summarize recent advancements in interactive machine learning by using merit-oriented taxonomy. We follow a similar approach like prior surveys but broadening our scope using merit-oriented taxonomy of iML.

In this paper, we review the state-of-the-art of interactive machine learning and thoroughly present research works in a merit-oriented taxonomy. While our main focus is on discussing iML contributions for robust, trustworthy and low resource machine learning, we have also presented issues related to iML performance evaluation. In section 2, we discuss about the overall merit-oriented architecture of iML. In section 3, adversarial attacks against black-box machine learning models are discussed. Moreover, it presents various iML inspired contributions aiming to combat adversarial attacks. We present researches conducted to increase the explainability and interpretability as a way to achieve trustworthy machine learning in section 4. Researches that employ human-in-the-loop for data processing and model building with the intention of lowering the data and computational resources are discussed in section 5. In section 6, we discuss the subjective nature of iML and related efforts to address problems in evaluating it. Further

analysis of findings and discussion is presented in 7. In the last section, a summary of the state-of-the-art review and potential research opportunities that are inspiring for future work on iML are discussed.

2 Merit-Oriented Taxonomy of iML

iML inspired contributions can be analysed from various perspectives. However, architectural Dudley and Kristensson [2018], Chatzimparmpas et al. [2020], Meza Martínez et al. [2019] and application/sector oriented Jiang et al. [2019], Liakos et al. [2018], Maadi et al. [2021] analysis has been significantly used techniques in the past. However, according to the meta-analysis Meza Martínez et al. [2019], Dudley and Kristensson [2018], Maadi et al. [2021], Liakos et al. [2018], Chatzimparmpas et al. [2020], Jiang et al. [2019], Amershi et al. [2014], Liu et al. [2017a], Teso and Hinz [2020] we conducted, most survey papers lack extensiveness and inclusiveness, leaving the various researches that constitute the state-of-the-art of iML untouched. Consequently, the employment of iML has been constrained to only predefined sectors or limited scope of the architecture.

To this end, we have thoroughly analysed the recent iML-inspired research works using merit-oriented taxonomy. After an extensive review of iML-inspired literature and untapped problems, we have categorized contributions into Robust Machine Learning Huang et al. [2011], Papernot et al. [2017], Carlini and Wagner [2017], Szegedy et al. [2013a], Sun et al. [2018], Brendel et al. [2017], Guo et al. [2019], Cheng et al. [2019a], Guo et al. [2018], Ma et al. [2019], Das et al. [2020], Slack et al. [2020], Emamjomeh-Zadeh and Kempe [2017], Trustworthy Machine Learning Teso and Kersting [2019], Gutzwiller and Reeder [2017], Krause et al. [2016], Ribeiro et al. [2016], Mozina [2018], Gutzwiller and Reeder [2017], Holzinger et al. [2018], Turchetta et al. [2019], Berkenkamp et al. [2016], Sui et al. [2018], Van Den Elzen and Van Wijk [2011], Liu et al. [2017b], Zhao et al. [2018], Mühlbacher et al. [2014], and Low Resource Machine Learning Ambati [2011], Frazier and Riedl [2019], Porter et al. [2013], Preuveneers et al. [2020], Holzinger et al. [2017], Fails and Olsen Jr [2003], Amershi et al. [2011], Tegen et al. [2020], Amershi et al. [2012], Dzyuba et al. [2014], Li et al. [2018], Jain et al. [2020], Fiebrink and Cook [2010], Gillian and Paradiso [2014], Schedel et al. [2011], Diaz et al. [2019], Arendt et al. [2018, 2017] based on their merit in the solution space. Besides, performance evaluation techniques that face bias because of the subjective nature of iML are analysed Boukhelifa et al. [2018], Krause et al. [2014], Ma et al. [2019], Yang et al. [2019], Krause et al. [2017], Felderer and Ramler [2021]. A high-level notional representation of merit-oriented architecture of iML is depicted in Figure 1. Our state-of-the-art review is channeled under these major merits of iML.
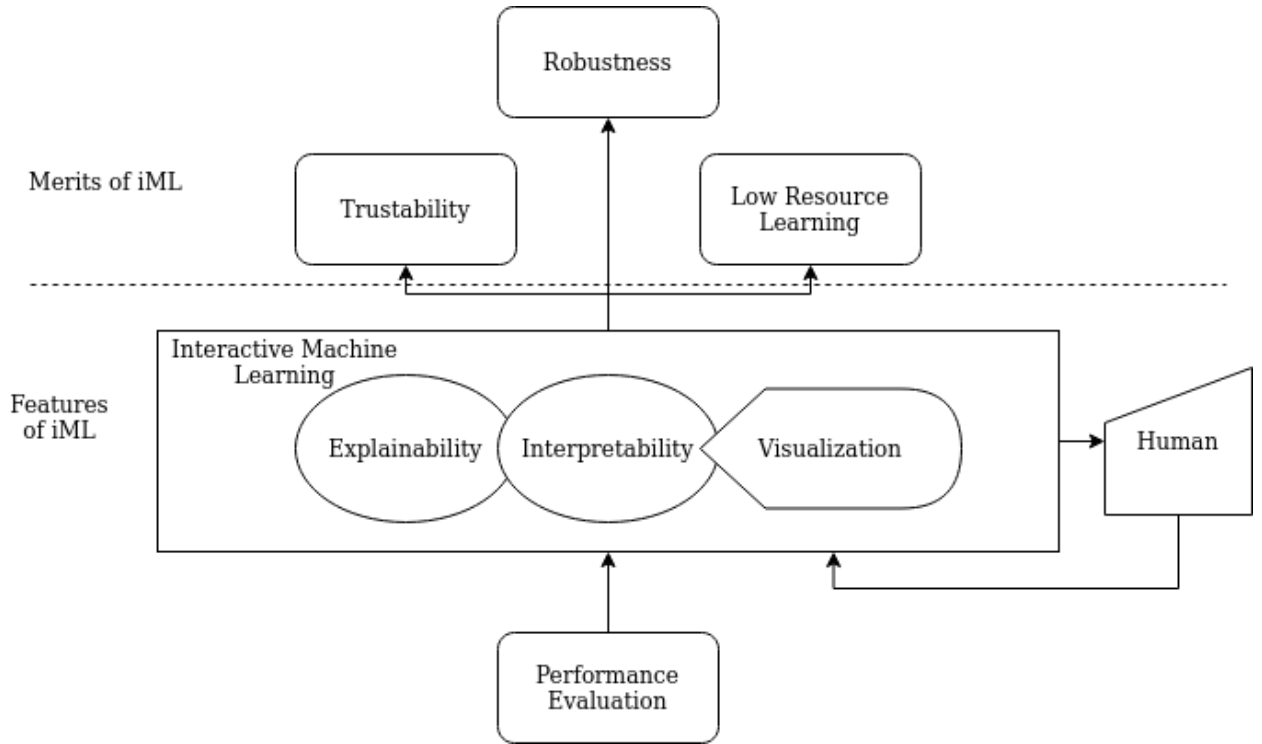

Figure 1: Merit-Oriented Architecture of Interactive Machine Learning

3   Robust Machine Learning

Deep learning algorithms are known for their vulnerability to adversarial attacks Szegedy et al. [2013b]. Adversaries can craftily manipulate legitimate inputs, which may be imperceptible to human eye, but can force a trained model to produce incorrect outputs. This issue is directly related to the black-box and intricacy nature of deep learning Papernot et al. [2017], Carlini and Wagner [2017], Szegedy et al. [2013a], Sun et al. [2018], Tu et al. [2019], Tramèr et al. [2017], Huang et al. [2011].

We use Papernot et al. [2016a] and Chakraborty et al. [2018] threat model to highlight the state-of-the-art of the adversarial attack. According to Papernot et al. [2016a], adversarial threat model is comprised of the following three dimensions, namely: the attack surface, adversarial capabilities, and adversarial goals. The attack surface refers to the overall data processing pipeline of machine learning from input to output and then action. Evasion attack Biggio et al. [2013], Kwon et al. [2018a,b], poisoning attack Chen et al. [2017], Zhang et al. [2019], Geiping et al. [2021] and exploratory attacks Papernot et al. [2017], Brendel et al. [2017], Guo et al. [2019] are the main scenarios considered under the attack surface. The aforementioned attacks can further be dissected into training phase Chen et al. [2017], Zhang et al. [2019], Geiping et al. [2021] and testing phase Biggio et al. [2013], Kwon et al. [2018a,b] attacks from the adversarial capabilities point of view. Data injection, data modification and logic corruption are among the strategies of training phase adversarial capability attacks. On the other hand, white-box Moosavi-Dezfooli et al. [2016], Tramèr et al. [2016] and black-box Papernot et al. [2016b], Rosenberg et al. [2017] attacks are among the testing phase capability attacks. Adversarial goal attack, on the other hand, infer adversary from the incorrectness of the target model. It is also further classified as confidence reduction, misclassification, targeted misclassification and source/target misclassification.

Adversarial attacks have real impacts on the robustness of a deep learning and other standard machine learning methods Wang [2021]. Therefore, exploring potential adversarial attacks and building a robust machine learning has been the focus of machine learning researchers.

There are various techniques being studied to address the problem of adversarial attacks. Researchers in the iML domain derived a number of strategies both to showcase the impact of adversaries Morris et al. [2020] and tackle Slack et al. [2020], Das et al. [2018] adversarial attack strategies discussed in Papernot et al. [2016a]. The fundamental assumption is that assuring the explainability and interpretability of black-box models by having the human-in-the-loop reduce the vulnerability of machine learning models to adversarial attack Rudin [2019].

iML has been used as a way to explain and explore model vulnerabilities to adversarial attacks as in Ma et al. [2019] and Das et al. [2020]. Specifically, Ma et al. [2019] enables exploration and explanation of model vulnerabilities to test-phase or poisoning adversarial attacks from the perspective of models, data instances, features, and local structures. A white-box exploratory attack based approach, Das et al. [2020], interactively visualizes neurons and their connections inside a DNN that are strongly activated or suppressed by an adversarial attack. Massif provides both a high-level, interpretable overview of the effect of an attack on a DNN, and a low-level, detailed description of the affected neurons.

In addition to explaining potentially perturbated inputs and models, iML based methods such as Slack et al. [2020] propose frameworks that masks the discriminatory biases of black-box classifiers. This plays a vital role to compensate the effects of perturbated inputs on a given model.

In addition to letting the user detect potential adversarial attacks and managing its impacts, iML has also been used to directly engage users in the model building process. This helps to avoid both training phase and test phase adversarial attacks as the user will be there validating inputs and intermediate results. A graph based framework, Emamjomeh-Zadeh and Kempe [2017], where user feedback is represented as edges and nodes are the models is a good example of such applications. It proposes Angluin's equivalence query model Angluin [1988, 1992] and Littlestone's online learning model Littlestone [1988] based on a robust machine learning framework that interactively learns models like classifiers Angluin [1988, 1992], orderings/rankings of items Joachims [2002], Radlinski and Joachims [2005], or clusterings of data points Awasthi et al. [2014], Emamjomeh-Zadeh and Kempe [2018], Balcan and Blum [2008]. In each iteration, the algorithm proposes a model, and the user either accepts it or reveals a specific mistake in the proposal. The feedback is correct only with probability $p > 1/2$ (and adversarially incorrect with probability $1 - p$), i.e., the algorithm must be able to learn in the presence of arbitrary noise.

Interactive machine learning has also been applied to tackle adversarial attacks on automatic speech recognition (ASR) models. For instance, Morris et al. [2020] designed a tool allowing experimentation with adversarial attacks and defenses on ASR. It implemented iML techniques and psychoacoustic principles to effectively eliminate targeted attacks. A similar and cloud-based tool to evaluate and compare state-of-the-art adversarial attacks and defenses for machine learning (ML) models is proposed in Das et al. [2019].

4   Trustworthy Machine Learning

To design and develop AI-based systems that users and the larger public can justifiably trust, one needs to understand how machine learning technologies impact trust Toreini et al. [2020]. The trustworthiness of AI-based systems is directly related to how the user is confident about the decisions made by the machine learning components. This may include its perception about both the intelligent model and knowledge Yin et al. [2019].

Users understanding of a given domain may negatively affect the trustworthiness of a model built around that specific domain. As it clearly stated in Honeycutt et al. [2020], the act of providing feedback can affect user understanding of an intelligent system and its accuracy. User trust toward a given model and even the perception of model accuracy might lower; regardless of whether the system accuracy is improved in response to their feedback. However, trustworthiness of a model should be defined considering its relative ability, benevolence and integrity.

To this end, we use the Trust Antecedent (TA) framework Mayer et al. [1995] definition of trust. According to this framework a trustee is given trust if it is perceived to have the ability, benevolence, and integrity toachive the desired goal. However, a black-box machine learning model that doesn't show its implementation details can not comply with the principles of this framework, hence affecting its trustworthiness. Besides, the black-box nature of these models significantly affects their applicability in in decision sensitive areas like health, finance, autonomous vehicles, criminal justice, etc.

Consequently, enhancing the trustworthiness of machine learning algorithms comes with making the model building process interactive Gutzwiller and Reeder [2017], Krause et al. [2016], Maadi et al. [2021]. Put succinctly, increasing the explainability and interpretability of machine learning by engaging user-in-the-loop increase its trustworthiness.

To this end, various explanatory frameworks that shows the implementation details of the model building process have been introduced. The essence of most exploratory algorithms is that intermediate results corresponding to a batch or even tuple are made subject to user-feedback whenever there is a variation between the predicted and the actual label.

In important contributions like Teso and Kersting [2019], Van Den Elzen and Van Wijk [2011], Mozina [2018], the iML is set to interactively query labels whenever the intermediate model fails to predict the label for each data point. For every prediction, the model provides the predicted label and explanations for the prediction. However, the way the the model and feedback are could be visual or textual.

Visual explainer systems are designed considering a set of guidelines proposed in various research contributions. For instance, Mühlbacher et al. [2014] propose a set of guidelines for integrating visualization into machine learning algorithms through a formalized description and comparison. Specific to automated iterative algorithms, which are widely used in model optimization, these researchers recommend exposing APIs for visualization developers. Using high-resolution APIs, visualization developers access the internal iterations for a tighter integration of the user in the decision loop.

Accordingly, Teso and Kersting [2019] propose a framework called CAIPI is proposed. It use a model-agnostic explainer, LIME Ribeiro et al. [2016], as fundamental component to compute the explainer and present them to the user as interpretable (visual) artifacts. Similarly, Van Den Elzen and Van Wijk [2011] propose a natural visual representation of decision tree structures where decision criterion is visualized in the tree nodes. BOOSTVis Liu et al. [2017b] and iForest Zhao et al. [2018] also focus on explaining tree ensemble models through the use of multiple coordinated views to help explain and explore decision paths.

Models and feedback from users can also be represented using textual-explanatory systems. An argument based explanatory iML algorithm Mozina [2018] propose a framework that exploits the usability of arguments to precisely axiomatize feedbacks both from the user and system side. It practically shows how to narrow the gap between domain experts and the machine learning model by engaging domain experts. Users provide feedbacks using a pair of reasons and outputs called arguments. Since arguments are generally presumptive and can't be untaken for a general set of predictions, the authors introduced prediction-level argument based explanations of decisions made by the learning model. Put succinctly, the training module generates initial features to perform prediction or classification. Whenever the learner notice major deviation between the desired and actual prediction, it consults the domain expert to provide feedback on the output (both the prediction and explanation). At this stage, the domain expert would be able to see the problem either in the predicted label or the rule yielding the outcome. Then, the user provides a set of arguments for each critical example (predictions with problem) and let the model to retrain keeping the feedback given from the user. Moreover, the paper dictated how the user should communicate with the machine learning component through a sequence of steps namely: Selection, Presentation, Argument formulation, Counter example detection, refinement, argument pruning. Another goal goal-oriented safe exploration algorithm that provably avoids unsafe decisions in real world problems is proposed in Turchetta et al. [2019].This framework takes suggested decisions as input and exploits regularity assumptions in terms of a Gaussian process prior in order to efficiently learn about their safety.

Exploratory machine learning has been applied to various decision sensitive sectors such as autonomous vehicles, health, and military. In Gutzwiller and Reeder [2017] used exploratory components to build trust among autonomous robots. Specifically, they analyzed the importance of using iML with neuro-evolutionary algorithms and conducted two separate tests were conducted on two autonomous environments; one with iML empowered ML algorithm and the other with black-box ML algorithms. The major problem they selected for this experiment is Search. Participants in this research engaged in building iML based search plan models. Finally, they were allowed to see and distinguish between the plans as iML generated or black-box generated. To this end, participants were able to choose iML based search plans for its effectiveness but with less trust relative to the black-box algorithm. The authors justified the preliminary knowledge of participants as the reason for their low degree of trust on the iML algorithm. In conclusion, the researchers convey the importance of using iML based neuro-evolutionary algorithms for searching problems of autonomous robots.

The application of iML in the health sector has also alleviated major computational trust related problems Holzinger et al. [2018], Berg et al. [2019]. A biological image data analysis tool called ilastik Berg et al. [2019] propose an iML framework that address the speed and usability requirements of machine learning by reasonably compromising model accuracy. Ilastik is able to formulate a feature space without the need to use bulk data as the other classical machine learning algorithms.

5   Low Resource Machine Learning

Low resource machine learning is a process of building an analytical model employing optimal resource utilization techniques. However, most machine learning algorithms have tradeoff between accuracy and resource utilization. As it is stated in Preuveneers et al. [2020], the most accurate model might be prohibitively expensive to computationally evaluate on a resource constrained environment. Consequently, the problem of building accurate and high performance machine learning models has been achieved at the expense of resource (data and computing) utilization. Nonetheless, sufficient amount of data and computing resources are not always at the stake. Besides, some problems may also require to be run on low resource setting. For instance, the use of pervasive devices and robots to build model is one valid scenario.

To this end, interactive machine learning help scientists and engineers exploit more specialized data within their deployed environment in less time, with greater accuracy and fewer costs Porter et al. [2013]. Various iML inspired researches has been conducted in the past couple of years. In this section, we will be discussing iML inspired solutions that aim at enabling low resource machine learning and improving the speed of models.

We have categorized our analysis of iML for low resource learning into small data machine learning and pervasive machine learning. In the small data machine learning section, we discuss contributions that use small data to build models in a low resource environment. On the other hand, the application of iML in pervasive low resource environment is discussed in the pervasive machine learning section.

5.1   Small Data Machine Learning

Machine learning algorithms usually requires a large volume of data in order to yield accurate result Holzinger et al. [2017]. However, big data is not always at stake to be used in some problem domains like under-developed languages, clinical trials, biomedical science and etc. Achieving small data machine learning requires optimal utilization of data. Engaging domain experts in the model building process in an interactive way would result in optimal utilization of important data items in a way contributing to the development of accurate models with small data.

To this end, interactive machine learning has a critical role. Hence, it has been studied by various researchers in the area of machine translation Ortiz-Martínez et al. [2010], Ambati [2011], González-Rubio et al. [2012], Peris et al. [2017], Lam et al. [2018], Foster et al. [1997], Huang et al. [2021], computer vision, search engine, social network analysis, music and games, and in data constrained sectors like health.

As it is mentioned above,iML based algorithms defines the state-of-the-art of machine translation systems. In such kind of systems, the knowledge of a human translator is combined with a MT system. For instance, Ambati [2011] propose a resource constrained machine translation. This paper explored active learning as an efficient way to reduce costs and make best use of human resource for building low-resource machine translation systems. Specifically, the author extended the traditional active learning approach of single annotation optimization to handle cases of multiple-type annotations. Furthermore, it show further reduction of costs in building low-resource machine translation systems. Reduction of the required annotated data inherently enabled low resource machine learning. Similarly, various classical IMT researches has been conducted in the few consecutive years. For instance, Ortiz-Martínez et al. [2010] propose an interactive online machine translation system that avoids the use of batch learning. In González-Rubio et al. [2012], an AL framework for interactive machine translation specially designed to process data streams with massive volumes

of data is presented. Recently, a couple of researches have also began to incorporate contemporary machine learning techniques for iMT. In Peris et al. [2017] and Lam et al. [2018], an interactive translation system, based on neural machine translation is deployed. Besides, a full-fledged iMT tools like Huang et al. [2021] has been proposed to further boost accuracy of machine translation systems.

Interactive machine learning also plays a significant role in image processing and computer vision in general.The use of iML for automatic feature selection is proposed in Fails and Olsen Jr [2003]. In this work, the researchers exploit the capacity of iML for automated feature selection. It show how to use iML to eliminate a manual feature selection. The training component of their architecture incorporates the feature selection sub component which further cooperates with the classifier, user feedback manager, and the user as a manual feedback provider. Furthermore, this research introduced an iML image processing framework called Crayons. Crayons is a toolto review and correct classifier errors by using customized decision tree as a learning algorithm and Mean Split Sub Sampled (MSSS) sampling technique.

Search engine optimization is also among the areas that has been studied by iML researchers. In a research contribution by

Besides, Fogarty et al. [2008], an interactive KNN based desktop application that enable re-rankings on a keyword-based web image search engine relying on the visual characteristics of images is proposed.

Another important iML enabled low resource learning is employed for pattern mining research problem. As it is known, pattern mining is an important process in exploratory data analysis. Various tools have been built to mine patterns from large datasets. However, the problem of identifying patterns that are genuinely interesting to a particular user remains challenging as it requires machine learning experts in addition to the domain expert Dzyuba et al. [2014]. To this end, iML plays a vital role both in enabling low resource pattern mining and assuring its valuability for specific users. For instance, Amershi et al. [2012] formulated a tool called ReGroup that provides users with filters that were generated based on features in the model. Put succinctly, ReGroup leverages the iML paradigm to assist users in creating custom contact groups in online social networks. The classifier provides a potential pattern of users (filters) in social network that likely qualifies for membership in a custom group that is being created. The users are set to provide feedback on each recommended filters where a selection of a contact for inclusion is taken as a positive sample and the remaining are skipped as negative samples. Likewise, the user refines the behaviour of the model with incremental improvements derived through iteratively applied user inputs. Participants noted that these filters provided insight in the patterns that were being exploited by the model, and thus served the dual purpose of explaining the model as well as their intended function as an interaction element. Besides, another research Dzyuba et al. [2014] use iML to formulate a pattern mining framework to mine user specific patterns. In this framework, the user is asked to rank small sets of patterns while a ranking function is inferred from this feedback by preference learning techniques. With a user supplemented by active learning heuristics, the resulting accurate rankings of patterns are used to mine new and more interesting patterns. They demonstrated their framework in frequent item-set and subgroup discovery pattern mining tasks. The ability of the framework to learn patterns accurately and the importance of its heuristic approach to ease users effort is presented precisely.

iML has also been applied to various game and entertainment research problems. A sketch-rnn, Ha and Eck [2017], present an interactive recurrent neural network (RNN) that help to construct stroke-based drawings of common objects. Besides their model encode existing sketches into a latent vector, and generate similar looking sketches conditioned on the latent space. Similarly, a research work in Jain et al. [2020] emphasize the design of a deep reinforcement learning agent that can play from feedback alone. This algorithm takes advantage of the structural characteristics of text-based games. Moreover, the application of iML in motion-driven music systems is presented in Fiebrink and Cook [2010], Gillian and Paradiso [2014], Schedel et al. [2011].Due to the complications to design a robust player-recognition or motion recognition system using standalone iML system, an enhanced model has been presented in Diaz et al. [2019]. They proposed an iML solution for Unity3D game engine in the form of a visual node system supporting classification (with k-nearest neighbour), regression (with multilayer perceptron neural networks) and time series analysis (with dynamic time warping) of sensor data.

The health sector is also known for data scarcity. In this sector, researchers are often confronted with only a small number of dataset or rare events, where a machine learning algorithm suffers from insufficient training samples Holzinger [2016], Maadi et al. [2021]. To this end iML researches has been conducted to build resource constrained model that adheres to the trustworthiness principles. In Holzinger et al. [2018], an interactive Ant Colony Optimization (ACO) technique is applied to the Traveling Salesman Problem (TSP). The TSP is an intransigent mathematical problem to find the shortest path through a set of points and returning to the origin. It is among the problems that resembles most computational problems in the health informatics sector. They introduce two novel concepts: Human-Interaction-Matrix (HIM) and Human-Impact-Factor (HIF) to control the ants action and the variable interpretation of the HIM respectively. The authors implemented their framework using a java-script based browser solution which has a great benefit of

platform independence and configuration. Although the results they found are promising, the researchers suggested further works like such as gamification and crowdsourcing to solve hard computational problems.

As we have discussed in the first section of this document, the essence of iML is to leverage the benefits of the human-in-the-loop. However, the presence of the human does not by itself guarantee the trustworthiness and performance of the model. The interaction between the user and the learning model should be supplemented with clear and in-depth visualization methods. To this end, Amershi et al. [2011] emphasize on how to design effective end-user interaction with interactive machine learning systems. They shape the human-model interaction in a way it answers these critical questions: which examples should a person provide to efficiently train the system, how should the system illustrate its current understanding and how can a person evaluate the quality of the system's current understanding in order to better guide it towards the desired behavior. Another practical contribution for human-model interaction is provided by Li et al. [2018]. They propose a new visual analytic approach to iML and visual data mining. Multi-dimensional data visualization techniques are employed to facilitate user interactions with the machine learning and mining process. This allows dynamic user feedback forms (data selection, data labeling and correction) to enhance the efficiency of model building. In particular, this approach can significantly reduce the amount of data required for training accurate models. Hence, it can be highly impactful for applications where large amount of data is hard to obtain. The proposed approach is tested on two application problems: the handwriting recognition (classification) problem and the human cognitive score prediction (regression) problem. Both experiments show that visualization supported iML and data mining can achieve the same accuracy as an automatic process can with much smaller training data sets. Furthermore, Arendt et al. [2018] propose a scalable client-server system for real-time iML framework called Computer-Human Interaction for Semi-Supervised Learning (CHISSL) Arendt et al. [2017]. The proposed system is capable of incorporating user feedback incrementally and immediately without a predefined prediction task. The light-weight computation web-client and heavyweight server constitute their architecture. The server relies on representation learning and off-the-shelf agglomerative clustering to find a dendrogram, which we use to quickly approximate distances in the representation space. The client, using only this dendrogram, incorporates user feedback via transduction. This work achieves low resource learning as it updates distances and predictions for each unlabeled instance incrementally and deterministically, with O(n) space and time complexity. Ultimately, this paper solved the scalability issue of CHISSL. Another important contribution in this regard is that of Talbot et al. [2009]. In this paper, EnsembleMatrix, SVM based visualization system, to tune classifier through user in the human-in-the-loop approach to build a superior model. This product presents a graphical view of confusion matrices to help users to directly interact with the visualization in order to explore and build combination models.

5.2 Pervasive Machine Learning

Interactive machine learning can also yield a better performance in pervasive computing environment where there are low computing resources. For instance, Frazier and Riedl [2019] present techniques to enable low resource computation for deep reinforcement learning agents complex behaviors in 3D virtual environments. Specifically, they considered an environment with high degree of aliasing, MineCraft, and conducted experiments with two reinforcement learning algorithms which enable human teachers to give advice-Feedback Arbitration, and Newtonian Action Advice under visual aliasing conditions. Similarly, Preuveneers et al. [2020] emphasizes on the importance of hyper tuning as a process to select the best performing machine learning model, its architecture and parameters for a given task. The fact that hyper parameter tuning does not take into consideration resource trade-offs when selecting the best model for deployment in smart environments has been taken as their research challenge. Consequently, the authors proposed a multi-objective optimization solution to find acceptable trade-offs between model accuracy and resource consumption to enable the deployment of machine learning models in resource constrained pervasive environments. On the other hand, Malloch et al. [2017] used pathward and fieldward techniques to enable creation of gestures that can be used by a touch-enabled pervasive mobile devices reliably. They introduced a novel dynamic guide that visualizes the negative space of possible gestures as the user interacts with the system.

On the other hand, Tegen et al. [2020] emphasises on the gaps of standard machine learning algorithms in dynamic sensor setting. Consequently, an iML based framework is implemented for "activity recognition" problems. This framework takes stream of data among different devices in the internet of things as an input. Three different machine learning approaches were used in the experiments: Support Vector Machine, k-Nearest Neighbor and Naıve Bayes classifier. These algorithms were customized to adapt the dynamic sensory setting. Uncertainty, Error (output from the model), State Change (what change in state urges the user to change label of data), Time (the time difference b/n queries and labeling by user), and Randomness (randomly assigning labels) are identified as factors that increase the subjectivity of users feedback. The results of this research work make it clear that the choice of interactive learning strategy has a significant effect on the performance when tested on recordings of streaming data.

6   iML Performance Evaluation

Although there are a number of invaluable contributions towards the goal of iML algorithms, the techniques used to evaluate their performance are not that mature. The standard machine learning algorithms use statistical model performance evaluation techniques which provide only a sole measure, obfuscating details about critical instances, failures and model features Ma et al. [2019]. However, the "human-in-the-loop" nature of iML systems triggers an extended need to diagnose the subjectivity of results. Therefore, iML performance evaluation should consider subjectively generated user evaluations and the cyclic nature of influence between the algorithm and users. This makes the evaluation of iML systems subjective and complex Boukhelifa et al. [2018], Yang et al. [2019].

Consequently, researchers proposed both customized and noble solutions to approach these critical issues. The iML Framework for Guided Visual Exploration (EvoGraphDice) Boukhelifa et al. [2018] is among the salient ones that is produced after an extensive experimental review of existing evaluation techniques. This paper founded its base mainly on the gaps of user based and system based feedback evaluation techniques to get insights on the underlying co-operation and co-adaptation mechanisms between the algorithm and the human. EvoGraphDice couples algorithm-centered and user-centered evaluations to bring forth insights on the underlying co-operation and co-adaptation mechanisms between the algorithm and the Human. Other papers focus on the input and output nature of iML models to diagnose their performance. For instance, the INFUSE system Krause et al. [2014] supports the interactive ranking of features based on feature selection algorithms and cross-validation performances. Another work Krause et al. [2017] also propose a performance diagnosis workflow. In this work, the instance-level diagnosis leverages measures of "local feature relevance" to guide the visual inspection of root causes that trigger misclassification.

Yang et al. [2019] identify diversified scenarios and subjective nature of explanations for the absence of benchmarks to evaluate explanations of iML algorithms.Consequently, they have defined the problem of evaluating explanations and systematically reviewed the existing efforts from state-of-the-arts. The authors discuss explanation as global (overall working structure) and local (particular model behaviour for individual instances) from scope perspective and as intrinsic (self-interpretable models) and posthoc (independent interpretation model) from dimension perspective. Furthermore, generalizability (generalization), fidelity (degree of exactness) and persuasability (comprehensibility) are identified as the three properties of explanation and corresponding methods has been revised for each aspect. As a result, they designed a unified evaluation framework based on the discussed properties of explanation and to the hierarchical needs from developers and end-users. In this work, three tiered explanation evaluation architecture is proposed corresponding to the three properties of explanation.

A similar work Felderer and Ramler [2021] discusses the importance of enhancing the quality of AI-based systems for a practical usage. They affirmed the importance of quality assurance to set benchmark for evaluations. Understandability and interpretability, defining expected outcomes as test oracles, and non-functional properties of AI-based systems of AI models are among the seven challenges they listed to assure quality of AI-based systems.

7   Discussion

An extensive merit-oriented state-of-the-art review of iML is presented in this paper. We used a bottom-up approach to categorize researches based on their primary importance. As shown in Table 1, robust machine learning, trustworthy machine learning, and low resource machine learning are identified as the major themes to categorize research works contributing to the state-of-the-art of iML. Significant issues related to iML evaluation has also urged us to consider it as one of the defining factors of the state-of-the-art of iML.

Furthermore, these research works have been further classified under their technical and sectoral emphasis. We subsume research works with specific component-wise and task-wise concentration under the technical section. Specifically, as shown in Table 1, HCI and Interface design, iML visualization, and model explanability related research works constitute the partial taxonomy of the technical section. Moreover, information processing class enclose research works focusing on the use of iML for training data management in areas such as machine translation. Searching and retrieval section contains iML inspired research works for search engine and retrieval systems. Adversarial attacks and iML inspired defense strategies are put under security and privacy section. Interactive machine learning inspired researches for low resource (pervasive) environment and clustering and optimization problems are classified under the technical section of the merit-oriented taxonomy accordingly. On the other hand, papers with special emphasis on general sectors like agriculture, health, game and development, and education are put under the sectoral section of the taxonomy. Finally, research works that focus on multiple classes and are generic to the topic we are studying are put under the other section of our iML taxonomy.

In the following sections, we discuss the major pillars of this state-of-the-art review. Moreover, the gaps identified in the revised research works are discussed as research opportunities.

Table 1. High Level Taxonomy of Interactive Machine Learning Researches

Merit-Orientation	Technical Section								Sectoral Section				Others
	HCI and Interface design	Visual Analytics	Searching and Retrieval	Security and Privacy	Information Processing	Model Explainability	Pervasive comp. and Robotics	Clustering and Optimization	Agriculture	Health	Education	Game and Entertainment	Others
Baselines and Meta Reviews	[2, 5, 31, 54, 83, 88]	[20, 64, 80]		[19, 97]	[111]	[116]			[78]	[54, 83]			[6, 36, 41, 119, 130]
Robustness	[26, 128]	[33, 82]	[39]	[14, 16, 18, 19, 21–23, 27, 28, 43, 49, 50, 60, 61, 72, 73, 89, 91, 95, 96, 106, 109, 113, 120, 121, 123, 135]	[1, 40, 48, 59, 76, 94, 98]	[107, 112]							
Trustworthiness	[13, 45, 47, 51, 57, 58, 71, 93, 131, 134]	[70, 81, 126, 136]		[67, 105, 124]	[25, 111]	[45, 92, 117, 136]	[12]	[9, 10, 81, 110, 133]		[11, 67, 103, 125]	[53, 127, 127, 137]		[86]
Low Resource Learning	[1, 3, 7, 8, 17, 24, 35, 35, 46, 56, 84, 99, 122]	[77, 114]	[39, 63, 65, 75, 104, 118, 129]	Security	[90, 100]	[38, 71]	[74, 85, 101, 115]	[9, 34, 42, 44, 55, 62, 65, 79, 87, 102, 123]	[29, 38]	[55]		[30, 37, 42, 52, 62, 66, 87, 108]	[4, 32]
Performance Evaluation	[15]	[68, 82]		[18, 132]	[69]								[129]

Figure 2: Merit-Oriented Architecture of Interactive Machine Learning

7.1 Robust Machine Learning

Tackling adversarial attacks has been one of the salient research areas in the areas of machine learning. Consequently, a number of adversarial example attacks have been discovered with the intention of exposing potential gaps on the learning models. However, such kind of approaches are reactive to solving robust machine learning.

From the attack surface perspective Papernot et al. [2016a], the nature of input perturbation Chen et al. [2017], Zhang et al. [2019], Geiping et al. [2021], Carlini and Wagner [2017], Szegedy et al. [2013a], test perturbation Biggio et al. [2013], Kwon et al. [2018a,b], and model exploitation Papernot et al. [2017], Brendel et al. [2017], Guo et al. [2019] has been studied extensively.

As it is thoroughly discussed from section 3, iML has been extensively applied to address the problem of model robustness. The methods against adversarial example attacks mainly basis on visualization of either the models intermediate output, data instances, local structures Ma et al. [2019], Emamjomeh-Zadeh and Kempe [2017] or the adversarial attacks Das et al. [2020], Morris et al. [2020], their masking techniques Slack et al. [2020] and visualization of the system's explanation Das et al. [2018].

Incorporation of an interactive component in the model building process is experimentally proved to have positive impact on tackling input perturbation. Human-in-the-loop can help to identify intentionally created parametric and data instance uncertainties. Similarly, users are observed to have successfully detecting either manipulated neurons (suppressed or hyper-activated) or data instances. Accordingly, the human-in-the-loop feature of iML evidently contributes in detecting adversarial attacks leading to the early enhancement of model robustness possible.

Given the increased complexity of adversarial attacks over-time and its subjective nature, the iML inspired research outcomes at hand falls short to sustainably address the problem. The power of model visualization and users knowledge towards advanced adversarial attacks should be studied to better solve the problem at hand.

7.2 Trustworthy Machine Learning

The most critical setback machine learning industry faces in recent years is a hampered applicability for decision sensitive sectors like Health, Finance, Military, and Autonomous Robotics like autonomous driving. Scientists hesitate to use standard machine learning for many trust-related reasons. The black-box nature of models, less control of users over data instances and absence of guidelines to formalize the integration of exploratory components in standard

machine learning are some of the reasons behind its hampered application. Besides, the relatively bigger error margin of standard machine learning algorithms and the small error sensitivity of models in the aforementioned sectors makes it hard to confidently employ standard machine learning algorithms.

To this end, we have reviewed research conducted to enhance trustworthiness of models. Put succinctly, we have reiterated on the contribution of iML algorithms in enhancing the trustworthiness of iML algorithms.

Most iML inspired techniques utilizes state-of-the-art textual Mozina [2018], Turchetta et al. [2019] and visualization components Van Den Elzen and Van Wijk [2011], Liu et al. [2017b], Zhao et al. [2018], Mühlbacher et al. [2014], Van Den Elzen and Van Wijk [2011], Ambati [2011], Teso and Kersting [2019], Liu et al. [2017b], Zhao et al. [2018].

The salient works in the textual explanatory aspects of iML reiterates mainly on deriving techniques to explain and interpret intermediate results of a model textually. The effectiveness of explanations mainly depends on their degree of interpretability by the users. Techniques like Argument based Axiomatization Mozina [2018] and goal-oriented safe-exploration Turchetta et al. [2019] has been reviewed as an important mechanisms to enhance the explanatory capability of models in general. On the other hand, visual exploratory methods are also found to provide better insights to complex models. Visualizations have been performed in either of intermediate results Teso and Kersting [2019] or the explanations themselves Liu et al. [2017b], Zhao et al. [2018]. Correlations between data, intermediate results and patterns among intermediate results have been used for visualization. Visualizations has also been applied to different data structures. We have seen techniques to visualize decision tree structures Van Den Elzen and Van Wijk [2011] and tree ensemble models through the use of multiple coordinated views Liu et al. [2017b], Zhao et al. [2018].

From the application point of view, it is demonstrated that iML inspired neuro-evolutionary algorithms boost trusts among autonomous agents Gutzwiller and Reeder [2017]. Similarly, iML can also be used to increase the trustability of models in the health sector Holzinger et al. [2018]. For instance, Kulesza et al. [2015] developed an explanation-centric approach to help end users to effectively and efficiently personalize machine learning systems. It intends to increase end users trust to a learning model by exposing the details of a learning model using an Explanatory debugging technique.

The other significant problem we have examined from the model trustworthiness perspective is absence of guidelines to build explanatory components. The employment of off-the-shelf interface design techniques resulted in poor performance of the overall algorithm Bernardo et al. [2017]. To this end, some works Mühlbacher et al. [2014] are conducted to set guidelines for integrating visualization into machine learning algorithms.

Besides, improving users experience can also boost its trustworthiness. In Bernardo et al. [2017], the importance of iML to improve user experiences by supporting user-centred design processes is studied. Besides, they conclude that there is a further role for user-centred design in improving interactive and classical machine learning systems.

From what we have revised, iML plays a vital role in introducing interactive learning and explainability.The fundamental assumption in this regard is that the more transparent model gains higher degree of trust. However, some researches have also tried to hypothesize an inverse relationship between trust and explainability Honeycutt et al. [2020]. Incidences like this might happen if users don't possess better understanding of black-box models and as a result focus on the uncovered error margins of transparent models. Similar findings reaffirm the importance of undergoing empirical model evaluations in addition to subjective ones in order to decide the trustworthiness of a given model.

Therefore, further researches on exploratory machine learning should be among the salient research problems in the iML domain. Besides, the design of user experience for iML should be studied further as it contributes for the effectiveness of the human-in-the-loop modality.

7.3   Low Resource Machine Learning

Machine learning consumes large data and resources in order to produce accurate models and undergone efficient performance evaluation. Generally, it is common knowledge that too little training data results in poor approximation. For instance, no human would need several thousand images to learn numbers in another language, yet a fully connected neural network needs too much data for good classification of MNIST dataset. On the contrary, too little test data will result in an optimistic and high variance estimation of model performance.

The amount of training data for machine learning has direct relationship with computing resources used to build models. The more data poured to the model builder, the more computational resource it requires. As a result, application of standard machine learning in resource constrained environments, more specifically in pervasive and robotics environments, has become less realistic. Sectors like under-developed linguistics, medical science, finance and military may not have sufficient amount of data for the standard machine learning framework to produce accurate outcome. Reducing the amount of data in these type of sectors compromises quality as the domain is sensitive to even smaller error margins.

Consequently, machine learning algorithms has been modified in a way it engages users-in-the-loop. Besides, it incorporates state-of-the-art visualization techniques to leverage users feed backs for data reduction and easing computing resource requirements. To this end, iML plays a vital role.

As we have clearly analysed in the state-of-the-art review, iML takes advantage of users feedbacks on intermediate results to reduce the amount of data and computational resources required to build well performing model. To this end, various iML inspired researches have been conducted. iML tools has been employed to enable low resource feature selection Fails and Olsen Jr [2003], Activity recognition in dynamic sensor setting Tegen et al. [2020], pattern mining and matching in social network analysis Dzyuba et al. [2014], data exploitation Amershi et al. [2012] and other applications. iML, especially its explanatory components, has been used to take advantage of some computing architectures. For instance, as we have precisely covered in this document, integration of iML to Deep Reinforcement Learning (DRL) has been experimented for 3D virtual environments Frazier and Riedl [2019]. This shows how well researchers recognized the importance of iML for resource constrained learning.

However, the contributions does not yet suffice the problems at hand. That means, iML is not exploited enough for critical demands of low resource learning especially in the area of robotics and pervasive computing. Robot simulation Marco et al. [2017], low resource object recognition and tracking, navigation and natural language capabilities of robots can be enhanced using iML inspired techniques. Therefore, exploiting iML for low resource machine learning especially in the area of machine intelligence and simulation is identified as a promising area of research.

7.4   iML Performance Evaluation

iML comes along with its complications of performance measure. Performance of iML inspired algorithms can not be effectively measured for two critical reasons namely: the subjectivity of users feedback and the co-operation and co-adaptation or the two sided influence between users and the system. The subjective nature of users in iML inspired model building makes it hard to put benchmarks for model evaluation Yang et al. [2019]. On the other hand, system based feedbacks are prone to ill interpretability. Consequently, the empirical performance measuring techniques of standard machine learning can not be exclusively used to assure the quality of iML inspired algorithms.

To this end some researches tried to address performance evaluation problems by addressing user's and systems feedback separately. Both the subjectivity of users feedback and degree of interpretability of systems explanation for a given intermediate results contributes to the complex nature of iML evaluation. Therefore, a partial effort to address one of these problems won't provide the desired solution. Similarly, focusing on user-centeric performance evaluation techniques does not suffice the problem at hand.

To this end, a coupled performance evaluation technique that utilizes both subjective and empirical performance evaluation techniques happens to yield better results Boukhelifa et al. [2018]. Interactive ranking of features Krause et al. [2014] and more specifically diagnosis of instances to leverage measures of local feature relevance are studied by different researchers. However, this area is still open for further research as there is no generalized and robust performance evaluation technique for iML as there is for statistical machine learning algorithms.

8   Conclusion and Future Works

We have seen how iML presents interesting challenges and prospects to conduct future research; not only in terms of designing robust algorithms and interaction techniques, but also in terms of coherent evaluation methodologies. Consequently, an extensive review of the state-of-the-art of iML is conducted in this paper.

From the tabular representation of iML inspired research works, shown in Table 1, it is clear that iML has not been exhaustively applied to areas such as searching and retrieval, pervasive computing and robotics, and clustering and optimization tasks. Moreover, although the importance of iML in decision-sensitive areas such as agriculture, health, education, game and entertainment is evidential, we have seen very few contributions in this regard.

Consequently, we have come to a conclusion that the domain of machine learning has not yet exploited the power of iML features and the human exhaustively. The merits of iML have not yet been fully enriched for contemporary machine learning algorithms. Identifying gaps and approaching them with state-of-the-art methods should be the primary goal iML researchers. To this end, we summarize potential research problems that we believe are inspiring for future work in iML as follows:

8.1 Adversarial Attack Defense

As per our analysis under the section of robust machine learning, the primary source of adversarial attacks is black-box nature of algorithms and bulk use of data. Interactive machine learning should further be studied in this aspect as it is evidently an efficient approach for small data and exploratory machine learning.

Further researches should be conducted to leverage the human-in-the-loop feature and suppress any training-phase and test-phase adversarial attacks. The engagement of the user in the model building pipeline will presumably reduce the degree of severity adversarial attacks against machine learning models.

8.2 Degree of User Engagement

The degree of users engagement in an interactive model building process is not well-defined. This may detrimentally affect model accuracy and overall performance as clearly presented in Honeycutt et al. [2020]. Therefore, a separate effort to set the degree of user engagement while building robust and trustworthy models should be researched thoroughly.

8.3 Exploratory Machine Learning

As we have discussed in the previous sections, empowering machine learning algorithms with exploratory features is a fundamental step towards transparent, trustworthy, and accurate model based computation.

However, providing explanations and interpretations on intermediate results of a model a critical problem until now. Not much work has been done in integrating iML features with contemporary machine learning algorithms. Therefore, we recommend further researches in building exploratory machine learning frameworks.

8.4 Low Resource Learning

Resource constrained learning is among the critical problems a machine learning industry face these days. The increased pervasiveness of computing nodes and the scarcity of data affected the applicability of machine learning significantly. Besides, the need for real time learning increased overtime. One important way to overcome this problem and ensure the applicability of machine learning in low resource machines is a systematic employment of iML. Especially, as we it is seen in our analysis, more effort on iML would enable real time learning on robots and other low resource agentsMordatch et al. [2015]. Furthermore, complex tasks like object detection, recognition and tracking, saliency identification, face and emotion recognition, natural language processing, robot navigation and simulation can be significantly enhanced with less data and computational resource by using interactive machine learning.

8.5 User Experience

The essence of interactive machine learning to engage users in the model building process and as a result obtain advantage of it. This can only be achieved if we provide a good user experience for the user subject to advise the learning model. To this end, user experience design plays a vital role.

Designing user interfaces using state-of-the-art design methods and in conformance with the specific requirements of iML algorithms Dudley and Kristensson [2018]. This allows experts outside the machine learning domain use machine learning algorithms to solve problems at their domain of expertise without much agony.

8.6 iML Performance Evaluation

Performance evaluation is one of the complications brought to the world of machine learning along with iML. The subjective, co-operation and co-adaptation nature of iML makes it hard to set benchmarks for performance evaluation. Therefore, we suggest further researching to enhance performance evaluation techniques for iML.

9 Acknowledgements

References

Dustin Arendt, Caner Komurlu, and Leslie M Blaha. Chissl: A human-machine collaboration space for unsupervised learning. In *International Conference on Augmented Cognition*, pages 429–448. Springer, 2017.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

Liu Jiang, Shixia Liu, and Changjian Chen. Recent research advances on interactive machine learning. *Journal of Visualization*, 22(2):401–417, 2019.

Andreas Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, Gloria Cerasela Crişan, C. Pintea, and V. Palade. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49: 2401–2414, 2018.

Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. Evaluation of interactive machine learning systems. In *Human and Machine Learning*, pages 341–360. Springer, 2018.

Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017a.

Jules Françoise. An overview of interactive machine learning, Oct 2020.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28, 2021.

Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

Mu-Huan Chung, Mark Chignell, Lu Wang, Alexandra Jovicic, and Abhay Raman. Interactive machine learning for data exfiltration detection: Active learning with human expertise. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 280–287. IEEE, 2020.

James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 29–38, 2008.

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013a.

Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. Iterated algorithmic bias in the interactive machine learning process of information filtering. In *KDIR*, pages 108–116, 2018.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019a.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

Reid Porter, James Theiler, and Don Hush. Interactive machine learning in data exploitation. *Computing in Science & Engineering*, 15(5):12–20, 2013.

Yuxin Ma, Tiankai Xie, Jundong Li, and Ross Maciejewski. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE transactions on visualization and computer graphics*, 26(1):1075–1085, 2019.

Nilaksh Das, Haekyu Park, Zijie J Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Chau. Massif: Interactive interpretation of adversarial attacks on deep learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2020.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

Ehsan Emamjomeh-Zadeh and David Kempe. A general framework for robust interactive learning. *arXiv preprint arXiv:1710.05422*, 2017.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

R. S. Gutzwiller and J. Reeder. Human interactive machine learning for trust in teams of autonomous robots. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 1–3, 2017. doi:10.1109/COGSIMA.2017.7929607.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5686–5697, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

M. Mozina. Arguments in interactive machine learning. *Informatica (Slovenia)*, 42, 2018.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration for interactive machine learning. *arXiv preprint arXiv:1910.13726*, 2019.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*, 2016.

Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4781–4789. PMLR, 2018.

Stef Van Den Elzen and Jarke J Van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 151–1inproceedings60. IEEE, 2011.

Shixia Liu, Jiannan Xiao, Junlin Liu, Xiting Wang, Jing Wu, and Jun Zhu. Visual diagnosis of tree boosting methods. *IEEE transactions on visualization and computer graphics*, 24(1):163–173, 2017b.

Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics*, 25(1):407–416, 2018.

T. Mühlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit. Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE Transactions on Visualization and Computer Graphics*, 20 (12):1643–1652, 2014. doi:10.1109/TVCG.2014.2346578.

Vamshi Ambati. *Active learning and crowdsourcing for machine translation in low resource scenarios*. PhD thesis, CMU- CS Depa, 2011.

Spencer Frazier and Mark Riedl. Improving deep reinforcement learning in minecraft with action advice. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 146–152, 2019.

Davy Preuveneers, Ilias Tsingenopoulos, and Wouter Joosen. Resource usage and performance trade-offs for machine learning models in smart environments. *Sensors*, 20(4):1176, 2020.

Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea, and Vasile Palade. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. *arXiv preprint arXiv:1708.01104*, 2017.

Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Effective end-user interaction with machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.

Agnes Tegen, Paul Davidsson, and Jan A Persson. Activity recognition through interactive machine learning in a dynamic sensor setting. *Personal and Ubiquitous Computing*, pages 1–14, 2020.

Saleema Amershi, James Fogarty, and Daniel Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30, 2012.

Vladimir Dzyuba, Matthijs van Leeuwen, Siegfried Nijssen, and Luc De Raedt. Interactive learning of pattern rankings. *International Journal on Artificial Intelligence Tools*, 23(06):1460026, 2014.

Huang Li, Shiaofen Fang, Snehasis Mukhopadhyay, Andrew J Saykin, and Li Shen. Interactive machine learning by visualization: A small data solution. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3513–3521. IEEE, 2018.

Vishal Jain, William Fedus, Hugo Larochelle, Doina Precup, and Marc G Bellemare. Algorithmic improvements for deep reinforcement learning applied to interactive fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4328–4336, 2020.

Rebecca Fiebrink and Perry R Cook. 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)*, volume 3, 2010.

Nicholas Gillian and Joseph A Paradiso. The gesture recognition toolkit. *The Journal of Machine Learning Research*, 15(1):3483–3487, 2014.

Margaret Schedel, Phoenix Perry, and Rebecca Fiebrink. Wekinating 000000swan: Using machine learning to create and control complex artistic systems. In *NIME*, pages 453–456. Citeseer, 2011.

Carlos Gonzalez Diaz, Phoenix Perry, and Rebecca Fiebrink. Interactive machine learning for more expressive game interactions. In *2019 IEEE Conference on Games (CoG)*, pages 1–2. IEEE, 2019.

Dustin Arendt, Emily Grace, and Svitlana Volkova. Interactive machine learning at scale with chissl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 819–824, 2012.

Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health*, 18(4):2121, 2021.

Ilker Kose, Mehmet Gokturk, and Kemal Kilic. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing*, 36:283–299, 2015.

P Qian, Y Zhou, and C Rudin. Using gaussian processes to monitor diabetes development. In *Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011) P. Qian, Y. Zhou, C. Rudin, eds*. Citeseer, 2011.

Ankita Tyagi, Ritika Mehra, and Aditya Saxena. Interactive thyroid disease prediction system using machine learning technique. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 689–693. IEEE, 2018.

André Dantas de Medeiros, Nayara Pereira Capobiango, José Maria da Silva, Laércio Junio da Silva, Clíssia Barboza da Silva, and Denise Cunha Fernandes dos Santos Dias. Interactive machine learning for soybean seed and seedling quality classification. *Scientific reports*, 10(1):1–10, 2020.

Simon Flutura, Andreas Seiderer, Tobias Huber, Katharina Weitz, Ilhan Aslan, Ruben Schlagowski, Elisabeth André, and Joachim Rathmann. Interactive machine learning and explainability in mobile classification of forest-aesthetics. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, 2020.

Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018.

Douglas J White. A survey of applications of markov decision processes. *Journal of the operational research society*, 44(11):1073–1096, 1993.

Marco Gillies. Understanding the role of interactive machine learning in movement interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(1):1–34, 2019.

Maria Kyrarini, Muhammad Abdul Haseeb, Danijela Ristić-Durrant, and Axel Gräser. Robot learning of industrial assembly task via human demonstrations. *Autonomous Robots*, 43(1):239–257, 2019.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *arXiv preprint arXiv:1906.06919*, 2019b.

Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

James D Hollan, Edwin L Hutchins, and Louis Weitzman. Steamer: An interactive inspectable simulation-based training system. *AI magazine*, 5(2):15–15, 1984.

Chenliang Zhou, Dominic Kuang, Jingru Liu, Hanbo Yang, Zijia Zhang, Alan Mackworth, and David Poole. Aispace2: an interactive visualization tool for learning and teaching artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13436–13443, 2020.

Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1283–1292, 2009.

Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

Ching-Chih Tsai, You-Zhu Chen, and Ching-Wen Liao. Interactive emotion recognition using support vector machine for human-robot interaction. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 407–412. IEEE, 2009.

Malihe Javidi, Baharak Shakeri Aski, Hale Homaei, and Hamid Reza Pourreza. A new approach for interactive image retrieval based on fuzzy feedback and support vector machine. In *2008 International Conference on Computational Intelligence for Modelling Control & Automation*, pages 1205–1210. IEEE, 2008.

Hee-Su Kim and Sung-Bae Cho. Application of interactive genetic algorithm to fashion design. *Engineering applications of artificial intelligence*, 13(6):635–644, 2000.

Chih-Chin Lai and Ying-Chuan Chen. A user-oriented image retrieval system based on interactive genetic algorithm. *IEEE transactions on instrumentation and measurement*, 60(10):3318–3325, 2011.

Sung-Bae Cho and Joo-Young Lee. A human-oriented image retrieval system using interactive genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 32(3):452–458, 2002.

Shang-Fei Wang, Xu-Fa Wang, and Jia Xue. An improved interactive genetic algorithm incorporating relevant feedback. In *2005 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2996–3001. IEEE, 2005.

E Poirson, J-F Petiot, and David Blumenthal. Interactive genetic algorithm to collect user perceptions. application to the design of stemmed glasses. In *Nature-Inspired Methods for Metaheuristics Optimization*, pages 35–51. Springer, 2020.

Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49(7):2401–2414, 2019.

Jingwen Meng, Xiaoming You, and Sheng Liu. Heterogeneous ant colony optimization based on adaptive interactive learning and non-zero-sum game. *europe PMC*, 2021.

Weikai Yang, Xiting Wang, Jie Lu, Wenwen Dou, and Shixia Liu. Interactive steering of hierarchical clustering. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

Fabricio Aparecido Breve. Simple interactive image segmentation using label propagation through knn graphs. *arXiv preprint arXiv:2002.05708*, 2020.

David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.

Angelos Chatzimparmpas, Rafael M Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020.

Miguel Angel Meza Martínez, Mario Nadj, and Alexander Maedche. Towards an integrative theoretical framework of interactive machine learning systems. *ais*, 2019.

Stefano Teso and Oliver Hinz. Challenges in interactive machine learning, 2020.

Josua Krause, Adam Perer, and Enrico Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics*, 20(12):1614–1623, 2014.

Fan Yang, Mengnan Du, and Xia Hu. Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831*, 2019.

Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 162–172. IEEE, 2017.

Michael Felderer and Rudolf Ramler. Quality assurance for ai-based systems: Overview and challenges (introduction to interactive session). In *International Conference on Software Quality*, pages 33–42. Springer, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013b.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 742–749, 2019.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016a.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon, and Daeseon Choi. Friend-safe evasion attack: An adversarial example that is correctly recognized by a friendly classifier. *computers & security*, 78:380–397, 2018a.

Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon, and Daeseon Choi. Multi-targeted adversarial example in evasion attack on deep neural network. *IEEE Access*, 6:46084–46096, 2018b.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Jiale Zhang, Junjun Chen, Di Wu, Bing Chen, and Shui Yu. Poisoning attack in federated learning using generative adversarial nets. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 374–380. IEEE, 2019.

Jonas Geiping, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller, and Tom Goldstein. What doesn't kill you makes you robust (er): Adversarial training against poisons and backdoors. *arXiv preprint arXiv:2102.13624*, 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016b.

Ishai Rosenberg, Asaf Shabtai, Lior Rokach, and Yuval Elovici. Generic black-box end-to-end attack against rnns and other api calls based malware classifiers. *arXiv preprint arXiv:1707.05970*, 282, 2017.

Jiakai Wang. Adversarial examples in physical world. In *Proc. Workshop Track Int. Conf. Learn. Represent.(ICLR)*, 2021.

John X Morris, Eli Lifland, Jin Yong Yoo, and Yanjun Qi. Textattack: A framework for adversarial attacks in natural language processing. *Proceedings of the 2020 EMNLP, Arvix*, 2020.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E Kounavis, and Duen Horng Chau. Adagio: Interactive experimentation with adversarial attack and defense for audio. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 677–681. Springer, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Dana Angluin. Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 351–369, 1992.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, 2005.

Pranjal Awasthi, Maria Balcan, and Konstantin Voevodski. Local algorithms for interactive clustering. In *International Conference on Machine Learning*, pages 550–558. PMLR, 2014.

Ehsan Emamjomeh-Zadeh and David Kempe. Adaptive hierarchical clustering using ordinal queries. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 415–429. SIAM, 2018.

Maria-Florina Balcan and Avrim Blum. Clustering with interactive feedback. In *International Conference on Algorithmic Learning Theory*, pages 316–328. Springer, 2008.

Nilaksh Das, Siwei Li, Chanil Jeon, Jinho Jung, Shang-Tse Chen, Carter Yagemann, Evan Downing, Haekyu Park, Evan Yang, Li Chen, et al. Mlsploit: A framework for interactive experimentation with adversarial machine learning research. In *Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, 2019.

Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 272–283, 2020.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

Donald Honeycutt, Mahsan Nourani, and Eric Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 63–72, 2020.

Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. Ilastik: interactive machine learning for (bio) image analysis. *Nature Methods*, 16(12):1226–1232, 2019.

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554, 2010.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, 2012.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220, 2017.

Tsz Kin Lam, Julia Kreutzer, and Stefan Riezler. A reinforcement learning approach to interactive-predictive neural machine translation. *arXiv preprint arXiv:1805.01553*, 2018.

George Foster, Pierre Isabelle, and Pierre Plamondon. Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194, 1997.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*, 2021.

Joseph Malloch, Carla F Griggio, Joanna McGrenere, and Wendy E Mackay. Fieldward and pathward: Dynamic guides for defining your own gestures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4266–4277, 2017.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

Francisco Bernardo, Michael Zbyszynski, Rebecca Fiebrink, and Mick Grierson. Interactive machine learning for end-user innovation. In *2017 AAAI Spring Symposium Series*, 2017.

Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P Schoellig, Andreas Krause, Stefan Schaal, and Sebastian Trimpe. Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with bayesian optimization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1563. IEEE, 2017.

Igor Mordatch, Kendall Lowrey, Galen Andrew, Zoran Popovic, and Emanuel V Todorov. Interactive control of diverse complex characters with neural networks. *Advances in Neural Information Processing Systems*, 28:3132–3140, 2015.