Towards a Taxonomy for Interpretable and Interactive Machine Learning

Elio Ventocilla[1], **Tove Helldin**[1], **Maria Riveiro**[1], **Juhee Bae**[1], **Veselka Boeva**[2], **Göran Falkman**[1], **Niklas Lavesson**[3]
[1] University of Skövde, School of Informatics
[2] Blekinge Institute of Technology, Department of Computer Science and Engineering
[3] Jönköping University, School of Engineering
{elio.ventocilla, tove.helldin, maria.riveiro, juhee.bae, goran.falkman}@his.se
veselka.boeva@bth.se
niklas.lavesson@ju.se

Abstract

We propose a taxonomy for classifying and describing papers which contribute to making Machine Learning (ML) techniques interactive and interpretable for users. The taxonomy is composed of six elements – Dataset, Optimizer, Model, Predictions, Evaluator and Goodness – where each can be made available for user interpretation and interaction. We give definitions to the terms interpretable and interactive in the context of user-oriented Machine Learning, describe the role of each of the elements in the taxonomy, and describe papers as seen through the lens of the proposed taxonomy.

1 Introduction

Due to the availability of large datasets, we enter a new era of *augmented intelligence*, where machines support humans to increase their cognitive capabilities. Indeed, when problems are complex and ill-defined, user-ML cooperation is needed. This approach to problem-solving is appealing for many reasons, for instance, to integrate valuable expert knowledge that may be hard to encode directly into computational models, to help resolve existing uncertainties as a result of error that may arise from automatic ML, or to build trust by making humans involved in the modeling or learning ML processes [Boukhelifa *et al.*, 2018]. A human and a machine collaborate to achieve a task, whether this is to classify objects, to find interesting data projections or patterns, or to design creative artworks [Boukhelifa *et al.*, 2018]. To date, several researchers have recently started to work at the intersection of human-computer interaction (HCI) and ML where the interaction with humans is seen as a central part of developing ML-systems.

The rapid increase of works related to interpretable and interactive machine learning (iiML) calls for an overview of this interdisciplinary subject, in order to structure the current literature and to develop a research agenda. One recent example of this type of study is presented in [Abdul *et al.*, 2018], focusing on providing an HCI agenda for explainable, accountable and intelligible systems. Another recent study [Lipton, 2016] focuses on interpretability in supervised learning and conducts a critical analysis of the literature to improve the specification of the task of interpretation. Interactive learning is explicitly put as out of scope in the paper.

In this paper, we unify the interactive and interpretable perspectives as we view them linked and in many cases interdependent. We tackle the challenge of classifying these works from an ML-component perspective, where we look at the components themselves and try to organize the studies found in the literature regarding the components that are being made interpretable or interactive, and how.

Thus, the main contribution of this paper is to present a taxonomy for categorizing the literature in the area of iiML. This taxonomy can be used to (1) provide a structured overview of the research work in the area of iiML, (2) identify research trends and opportunities for research in the future and (3) suggest a standard terminology for iiML.

The paper describes, first, relevant background and terminology in the context of iiML (see section 2). The method followed for outlining the taxonomy is summarized in section 3. The taxonomy and its components are described thereafter in section 4. We discuss implications in section 5 and conclude the paper with some conclusions and final remarks in section 6.

2 Background concepts

The use of ML-based support systems has shown great potential in multiple areas such as education, healthcare, manufacturing, retail, etc. To fully exploit the benefits of such systems in our daily activities, we need to make these technologies more accessible to all. Nevertheless, many conventional applications of ML are mostly agnostic to the fact that their inputs and outputs are directed at humans [Amershi *et al.*, 2013]. To resolve this, researchers from the areas of HCI and AI now collaborate, trying to make these systems more transparent and understandable, providing explanations, visualizing the inner workings of these complex ML systems or supporting human interaction with the core elements of ML.

Several concepts and terms have arisen in this quest, for instance, *interpretability*, *interactivity*, *understandability*, *explainability*, *comprehensibility*, *intelligibility* and *transparency*. This section gives a very brief summary of some definitions found in the literature for these terms, in order to provide an introduction of the rich terminology used by researchers within the iiML area.

Interpretation in relation to ML techniques is defined by Chuang *et al.* [2012] as the "facility with which an analyst makes inferences about the underlying data". The term *transparent*, on the other hand, is related to ML techniques that, (1) produce models that a typical real-world user can read and understand; (2) use algorithms that a typical real-world user can understand, and (3) allow a real-world user to adapt models to new domains [Chiticariu *et al.*, 2015].

Comprehensibility, *understandability* and *interpretability* are regarded as synonyms in [Piltaver *et al.*, 2016]. Comprehensibility is defined by Zhou [2005] as a property of ML algorithms that "produce patterns understandable to human beings". In the same paper the author refers to a postulate made by Michalski [1983], which states that the resulting elements of a computer induction "should be comprehensible as single 'chunks' of information, directly interpretable in natural language". Moreover, the term *explainable* is found to be used in conjunction with the terms intelligibility [Abdul *et al.*, 2018], understandable [Stumpf *et al.*, 2009] and transparent [Lim *et al.*, 2009].

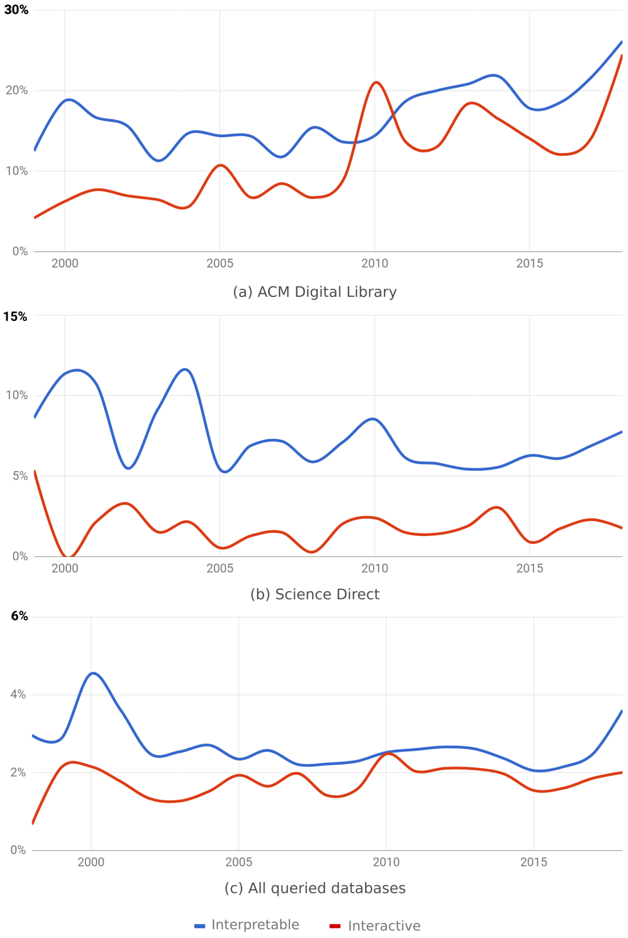
Most often the terms above are used to describe ML components that readily lend themselves, or are presented in a way, which is understandable to humans. For a closer discussion of the terms, please refer to [Bibal and Frénay, 2016].

The term *interactive* is, in this context, referred by Fails and Olsen [2003] as a property of models which allow users to "train, classify/view and correct the classifications". Holzinger *et al.* [2016] expand the interaction boundaries from users to agents, where agents can also be other systems. Fiebrink *et al.* [2011] highlight the importance of model evaluation for effective user interaction in model improvement. A term related to interactive ML often found in the literature is that of human-in-the-loop (e.g. [Holzinger *et al.*, 2016; Bohanec *et al.*, 2017; Lee *et al.*, 2017]), and is used to describe a role played by users in improving the "goodness" of a model by giving feedback to the ML algorithm, and/or "steering" its computation.

In our work, we use interpretable and interactive ML as an overall term that comprises all of the above concepts, where focus is put on placing the human user in the center of the ML processes. As we see it, interpretability is enabled through an ML system's explainability, transparency, intelligibility and understandability, and is often realized through the user interacting with the ML system, enabling the user to obtain a better understanding of the system's inner workings as well as to improve its output.

3 Method

The taxonomy has been developed through the following process: (1) *keyword and query definition*, i.e., establishing search keywords with which papers were queried in the databases; (2) *paper ranking*, i.e., sorting and choosing papers based on a given criteria; (3) *paper survey*, i.e., individually surveying the contents of the papers based on their relation to ML interaction and interpretation; (4) and *discussion*, i.e., coming to an agreement between all authors about common elements found in step (3). The last two steps, *paper survey* and *discussion*, were iterated as often as needed for all authors to come to a consensus on the common elements found and their relations.



(a) ACM Digital Library

(b) Science Direct

(c) All queried databases

Figure 1: Trend over the past 20 years of ML journal and conference publications matching our queries on interpretable and interactive ML. These represent results from (a) ACM Digital Library, (b) Science Direct, and (c) all queried databases, i.e., ACM Digital Library, Science Direct, IEEE Xplore, Scopus, Web of Science, NIPS and MLR. The y-axis depicts percentage, that is, the number of papers related to interpretable or interactive ML divided by the total amount of ML papers for the given year.

Two queries were used in step (2), one for interpretable ML and another for interactive ML. The query for interpretable ML was as follows:

```
  TAK=(``machine learning'' AND
(interpretable OR understandable OR
comprehensible OR explainable OR
intelligible OR transparent))
```

Where TAK stands for paper Title, Abstract or Keywords. The query for interactive ML was the following:

```
  TAK=(``machine learning'' AND
(interactive OR ``human-in-the-loop''))
```

The query keywords were agreed upon by all authors and were then used in the following databases: the ACM Digital Library, IEEE Xplore, Scopus, Science Direct and Web of Science. Figure 1 shows the results to these queries. Journals and conference papers from each database were sorted by relevance (as given by each database) and only the first 200 papers of each database were taken. Additionally, we crawled two web pages, NIPS[1] and PMLR[2], and ran the queries on the retrieved titles and abstracts – these web pages do not provide paper keywords. The resulting papers were ranked by their average number of citations per year and then reduced to only those with two or more average citations. Figure 2 shows a distribution of papers following this restriction: a total of 357 papers with 208 matching interpretable ML and 149 interactive.

From the total set of ranked papers, twenty six papers were surveyed by all authors in an individual manner, as an exercise for the development of the taxonomy. These papers were picked subjectively within the ranked list.

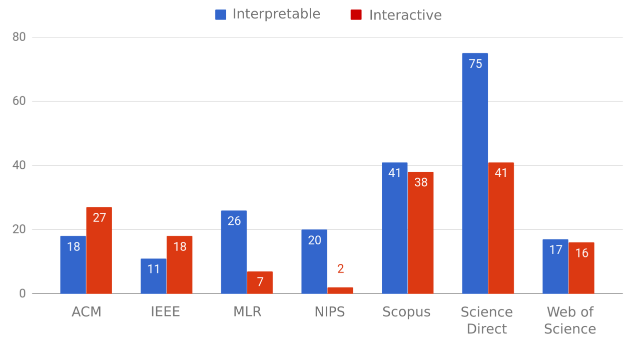


Figure 2: Number of papers per database, matching the queries for interpretable and interactive ML, with average citations per year equal or higher than two.

4 Taxonomy

The proposed taxonomy (Figure 3) is composed of six elements: Dataset, Optimizer, Model, Predictions, Evaluator and Goodness. Three of the elements – Optimizer, Model, and Evaluator – are based on Domingos's [2012] components of ML, whereas the other three – Dataset, Predictions and Goodness – have been added based on our observations from the surveyed papers. These observations are in line with the results reported in [Glauner *et al.*, 2017], where it is noted that a broader view on machine learning is needed, which includes not only the model but also the data, optimization techniques and evaluation metrics, a view that has, so far, been largely ignored in the literature. In Figure 3, white arrows represent inputs and black arrows outputs. Black boxes represent elements which produce an output given one or more inputs. The optimizer component (O), for example, has two inputs, a training dataset (X) and a model (M), and one output which is a new optimized model (M).

Each of the components in the taxonomy can potentially be enabled for user interpretation and/or interaction, as shown in the following subsections. The papers selected are not exclusive in the sense that they can only be used as illustrations for one of all the components shown in Figure 3, but can indeed showcase interpretability and interactivity within several of the components. Thus, the purpose of the examples is to describe in more details our view upon the taxonomy and its classifications.

4.1 Dataset

Dataset (X) regards training data, test data, validation data and/or input data for prediction or classification. It is expected to interpret and interact with the selection of parameters/features, its predictive probabilities, and the quality of the classified/predicted outcomes. It ultimately needs to support the users to make the next decision.

A solution which contributes to user **interpretation** of a dataset is performed by step by step user changes to the original feature values and allowing the users to compare/rank the models or classification/prediction results. The user needs to have the overview of the parameters selected as well as the corresponding results since following up the incremental effects is crucial. In fact, interpretation of a dataset is very closely connected to user interaction. A solution which contributes to user **interaction** with a dataset is one that lets the user have control of the penalizing/rewarding activities or which enables the user to simply change a value (e.g., feature weight) to see the corresponding classification results or predictions, mostly in real time. It is very related to user interpretation which guides the user in the model creation/update by interacting with the input data.

Two examples of research papers that contribute to the interaction with interpretation of predictions or classifications are [Krause *et al.*, 2016] and [Krause *et al.*, 2014].

In [Krause *et al.*, 2016], a tool is presented which detects diabetes and makes risk predictions by testing the patients' glucose measures. By tweaking features, they find the most impacting feature that brings a high-risk level of diabetes. The tool allows users to interpret how features affect the prediction with the help of visual representations to interact and compare. The interaction with the tool helps the user to diagnose the dependencies of features and to find the feature with the most impact.

[Krause *et al.*, 2014] bring insights to clinical researchers predicting patient outcomes by manipulating factors that affect a disease e.g., diabetes. They developed a visual analytics tool that enables interactive feature selection on high dimensional data with the ranking results of multiple feature selections, cross-validation folds, and classifiers. The predictive power is evaluated by ranking features (feature selection algorithms: information gain, Fisher score, odds ratio, relative risk) across multiple classification algorithms (tree, logistic regression, naive Bayesian, k-nearest neighbors) for the users to see the relevant features and compare the results.

4.2 Optimizer

The optimizer (O) – or optimization algorithm – is in charge of improving a model depending on the ML problem (e.g., classification, regression, clustering). To do so, it takes two inputs, a training data set X and a model M, and constructs an

[1] https://papers.nips.cc/ [Accessed 2018-05-15]

[2] http://proceedings.mlr.press/ [Accessed 2018-05-15]

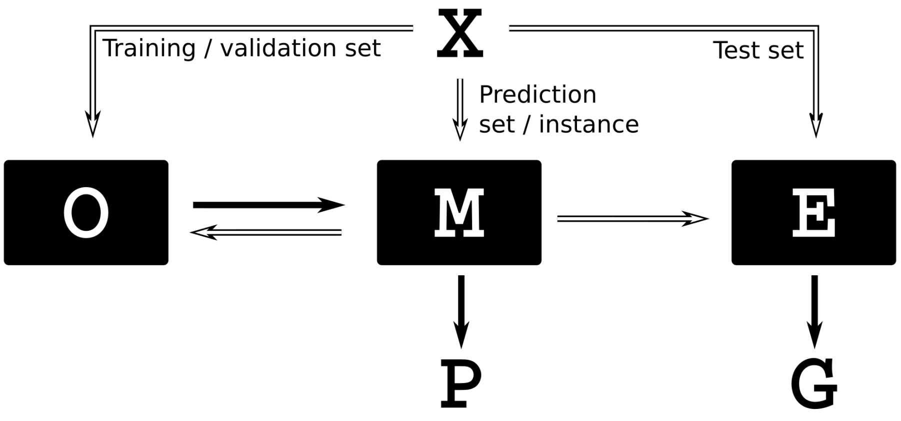


Figure 3: A taxonomy for interpretable and interactive Machine Learning. 'O' stands for optimizer, 'M' for model, 'E' for evaluator, 'P' for prediction, 'X' represents a dataset and 'G' the goodness of the model. Arrows depict inputs (white arrows) and outputs (black arrows). Each component can be potentially "opened" for user interpretation and interaction.

improved instance of M that better describes or generalizes X. Examples of optimizers are Linear Regression and Quasi-Newton methods such as DFP and BFGS.

Note that our definition is that of a component and not of a process. In that sense, an optimizer is not an optimization or a learning process. An optimizer is, however, an element of such a process. The distinction is important because elements of a process, and the interactions they can provide, are different. These processes are relevant to iiML but are outside of the scope of this paper. We dedicate, nevertheless, some space to the topic in Section 5.

A solution which contributes to user **interpretation** of an optimizer is one which discloses, in a *simple* human-readable manner, optimizer parameters (e.g., the learning rate or a distance function) and their impact to the output (i.e., the optimized model). A solution which contributes to optimizer **interaction** is one which allows users to change views of the optimizer, zoom in into its details, *tune* its parameters (e.g., change the learning rate or the distance function), or even change the optimizer itself.

Two examples of research papers that contribute to the interaction with, and interpretation of, an optimizer are [Holzinger *et al.*, 2016] and [Schreck *et al.*, 2009]. The former describes how a user can see and influence the behavior of the Ant Colony Optimization algorithm in the context of the Traveling Salesman Problem. Interpretation is facilitated through visual cues in a graph, with edges representing paths and their width the level of pheromones; interaction, on the other hand, is provided by allowing the user to manipulate pheromone levels. Schreck *et al.* [2009] describe a system which allows users to change the learning rate, as well as the neighborhood kernel, of a Self-Organizing Map for clustering trajectories. Such updates are then reflected on the new optimized versions of the model in the forthcoming iterations of the learning process.

4.3 Model

Learning algorithms are used to create a mathematical abstraction or generalization of data. This abstraction is called a model – or representation [Domingos, 2012] – and is represented by the 'M' box in Figure 3. Models are often implemented in a way that, given a new observation X, produces a classification or prediction P, e.g., given the profile of a client X, compute the risk of, e.g., giving him/her a bank credit (a classification P in the form of *low risk* - *high risk*).

A solution which contributes to user **interpretation** of a model is one which produces or wraps the mathematical abstraction – i.e., parameters, expressions, structure – in a format which facilitates human inspection and a human-readable explanation of its logic. Moreover, a solution which contributes to model **interaction** is one which allows users to change views of the model, get details on demand, and/or manipulate its inner elements (e.g., change parameters or its structure). Such solutions should support transparent interaction with humans without requiring that a user has expert knowledge of the ML techniques used. For example, a solution which computes and communicates ML results in ways that are compatible with the human decision-making process, and that can readily incorporate human experts' domain knowledge can be said to be interpretable and interactive.

Two examples of research papers that contribute to the interaction with, and interpretation of, a model are [Letham *et al.*, 2015] and [Hu *et al.*, 2014]. Letham *et al.* [2015] present a generative model called Bayesian Rule List which produces models in the form of small sets of *if... then* rules. Their contribution claims a balance between interpretability, accuracy and computational demand. Hu *et al.* [2014] developed a framework and a system for integrating user feedback into topic models in an interactive manner. After modeling a given number of topics 'M', their system allows users to modify them by adding, removing or increasing the relevance of words.

4.4 Prediction

The prediction component (P) of the taxonomy proposed regards the explanation of a prediction or classification produced by a model (M) to the human user as well as the possibility for the user refine it.

A solution which contributes to user **interpretation** of predictions or classifications is one which *explains* these result to the human user, i.e., why has X been classified as Y (and not Z)? As the actual features of the model can be difficult for a non-ML expert to interpret, the explanations generated need sometimes include other features than those used by the model for ensuring efficient human interpretation. A solution which contributes to user **interaction** with predictions or classifications is one which enables the user to investigate and tune how different parameter settings affect the resulting prediction/classification output, thus setting a foundation for increased knowledge of the workings of the model.

Two examples of research papers that contribute to the interaction with, and interpretation of, predictions or classifications are [Ribeiro *et al.*, 2016] and [Kulesza *et al.*, 2015]. In [Ribeiro *et al.*, 2016] an explanation technique is presented which explains the prediction of a classifier in an interpretable manner. For example, if applied in a medical scenario, the user is presented with a classification of the patient's illness/status together with the evidence for/against this classification. By inspecting these evidence, the users can use their expert knowledge of the domain to determine whether to trust the classification or not. [Ribeiro *et al.*, 2016] further argue that the explanations also can be used to select the most appropriate model for the problem at hand, by comparing the predictions of several models with the user's expert knowledge.

In [Kulesza *et al.*, 2015], an approach towards explainable ML is suggested, where the users are presented with explanations for the system's predictions to enable them to build mental models of the learning system as well as to interactively personalize it. The features of the classifier that are used to make the prediction are presented to the user, together with how each feature contributed to the prediction. Font size and color is used to convey this information, together with percentages of the likelihood of the prediction being correct. To correct the predictions made, the user is allowed to input or remove features from the explanation, which in turn will add or remove those features from the ML model's feature set. The users are also enabled to adjust the importance of the features in the explanation by increasing/decreasing the size of the feature in the interface, resulting in a higher/lower weight of the feature in the learning model.

4.5 Evaluator and Goodness

The evaluator (E) carries out the assessment of the performance of a model (M). Typically, traditional objective metrics from ML and Data Mining are used for this purpose, for instance, accuracy, precision, recall, squared error, f-score, information gain, etc. Such metrics are typically specialized to the type of machine learning problem and method used, i.e., clustering, classification, regression, etc. The input of the evaluator is usually the model itself and a test set, in order to assess the performance or "goodness" (G, output) of the model.

The evaluation of the model might be different from the overall performance of the ML-based system. Since we are considering interpretable and interactive ML solutions, the overall evaluation, and also the internal one, might include subjective metrics as well, commonly used in HCI, for example, usability evaluations (how long did users take to carry out certain tasks, were they successful, how many errors did they make, how many commands/features did they use, etc.). An example of model evaluation that includes subjective assessments is, for instance, presented in Amershi *et al.* [2010]. General challenges related to model evaluation are discussed in Fiebrink *et al.* [2011], where the authors conclude, among other issues, that exploratory evaluations of models can complement objective metrics in allowing users to evaluate models against a wide range of criteria.

A solution which contributes to user **interpretation** of an evaluator is one which allows the user to understand the evaluation, for example, showing the results of the performance of the predictions through visualizing the accuracy, error rate, etc. A solution which contributes to user **interaction** with an evaluator is one which supports user understanding and tuning of the evaluation process.

Two examples of research papers that contribute to the interaction with, and interpretation of an evaluator are [Bohanec *et al.*, 2017] and [Kapoor *et al.*, 2010], respectively. Bohanec *et al.* [2017] present a framework for explaining the results of classification models. According to our taxonomy, the evaluation carried out by Bohanec *et al.* is interpretable, since the explanations provided are claimed to support a better understanding of the classification accuracy of the models. Kapoor *et al.* [2010] present *ManiMatrix*, a system that support users in classification tasks using ML. The visualization of intermediate steps of the process supports and enhance the classification process, in some cases outperforming the highest automatic accuracy ever published for the problems in question. The evaluation presented in [Kapoor *et al.*, 2010] uses an interactive confusion matrix, which represents classification results by aggregating instances within a grid; each row in the matrix represents an instance's true class and each column an instance's predicted class. The users can specify interactively an increase or decrease in the tolerance for numbers of cases classified into each cell.

5 Discussion

The proposed taxonomy gives a detailed view of ML components and works as a reference for structuring how users can interpret and interact with each of them. The examples given illustrate some of the different ways of how the research community has contributed to iiML.

The taxonomy represents low-level components of ML. It is low-level for it does not explicitly depict higher processes such as, e.g., the machine learning process or – at an even higher level – the decision-making process. Such a fine-grained taxonomy can be challenging to use, for it requires a deeper understanding of how ML systems are implemented. Some system implementations have layers in between the user and the ML components, as means to map and transform input from one end to the other. Additionally, a user interaction can, in cases, trigger chains of transformations across all ML components. Telling where the impact of the user feedback will take effect is not always straightforward.

The research community has contributed to user involvement in higher level processes such as the algorithmic learning process. Such a process can be found implicitly in the *Optimizer-Model* loop. Human interpretation at this level can involve process-wise elements such as [Mühlbacher *et al.*, 2014]: aliveness, i.e., status of the learning process (e.g., "learning in progress" or "has failed"); and progress, e.g., estimated remaining time. Human interaction, on the other hand, may translate to user control over the process, e.g., cancel execution, prioritize work [Mühlbacher *et al.*, 2014]. Contributions on this level, and others, are relevant to iiML but are not in the scope of this paper.

We envision research challenges at the current stage of our work. Challenges with the dataset component (X) may be given by the complexity of the data itself. Even relatively small datasets can be very complex to understand and handle. Interacting and interpreting graph or image data will probably prove more difficult than tabular numeric data. A challenge with optimizers (O) is their disassociation to the knowledge domain of the task, that is, they reside in a mathematical realm to which users might not relate to. The same might be said about models (M), although in their case certain formats have proved to be more interpretable than others (e.g. decision trees in contrast to neural networks). A recurrent challenge with models is the trade-off between interpretability and accuracy.

Papers reviewed in this domain that present evaluations (E) use either traditional objective performance methods and metrics from ML/DM or subjective assessments from HCI. Few present a combination of both strategies. We think that there is a lack of methods and metrics that can assess the overall performance of human-machine collaboration, which go beyond the evaluation of specific components of the whole system. Therefore, evaluation and metrics that combine both strategies are needed in the future.

6 Conclusions and Future Work

This paper presented a taxonomy for classifying and describing papers in the area of iiML. The aim of such taxonomy is to structure the literature found in this interdisciplinary area up to now and investigate through examples how ML can become more interpretable and interactive. The proposed taxonomy has six components, Dataset, Optimizer, Model, Prediction, Evaluator, and Goodness. We provided a description of each along with relevant papers to illustrate their role in iiML.

We provided brief descriptions of the different terminologies used under the scope of iiML, but believe there is a need for an agreed upon terminology to be used by HCI and ML practitioners, in order to better structure future work within the area. By exemplifying how the human user can be incorporated into the various ML components, we hope that our work can inspire practitioners within the field to develop highly functioning iiML systems where the strengths of the humans and the machines are efficiently exploited.

Through our work, we have identified an increasing interest of iiML as a research field, yet also the lack of examples of ML systems where the user is incorporated in all of the ML components outlined. With this paper, where we have identified low-level components on iiML, we hope to highlight this challenge and the need for addressing it in future work within the field.

As future work, we intend to investigate and describe higher-level processes of iiML, such as the learning process. A better understanding of how these processes are implemented, and how they are tailored for user interpretation and interaction, can prove useful for better structuring current research and for outlining feature research. We expect the findings to be helpful for building a complete taxonomy for iiML.

Acknowledgments

We would like to thank Nikolas Huhnstock and Niclas Ståhl at the Univeristy of Skövde for fruitful discussions regarding the proposed taxonomy.

References

[Abdul *et al.*, 2018] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 582. ACM, 2018.

[Amershi *et al.*, 2010] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Examining multiple potential models in end-user interactive concept learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1357–1360. ACM, 2010.

[Amershi *et al.*, 2013] Saleema Amershi, Maya Cakmak, W Bradley Knox, Todd Kulesza, and Tessa Lau. IUI workshop on interactive machine learning. In *Proceedings of the Companion Publication of the International Conference on Intelligent User Interfaces*, pages 121–124. ACM, 2013.

[Bibal and Frénay, 2016] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: An introduction. In *Proceedings on ESANN*, pages 77–82, 2016.

[Bohanec *et al.*, 2017] Marko Bohanec, Marko Robnik-Šikonja, and Mirjana Kljajić Borštnar. Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management & Data Systems*, 117(7):1389–1406, 2017.

[Boukhelifa *et al.*, 2018] Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. Evaluation of interactive machine learning systems. *arXiv preprint arXiv:1801.07964*, 2018.

[Chiticariu *et al.*, 2015] Laura Chiticariu, Yunyao Li, and Fred Reiss. Transparent machine learning for information extraction. *EMNLP (tutorial)*, 2015.

[Chuang *et al.*, 2012] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2012.

[Domingos, 2012] Pedro Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87, October 2012.

[Fails and Olsen, 2003] Jerry Alan Fails and Dan R. Olsen, Jr. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 39–45, New York, NY, USA, 2003. ACM.

[Fiebrink *et al.*, 2011] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 147–156, New York, NY, USA, 2011. ACM.

[Glauner *et al.*, 2017] P. Glauner, M. Du, V. Paraschiv, A. Boytsov, I. Lopez Andrade, J. Meira, P. Valtchev, and R. State. The top 10 topics in machine learning revisited: A quantitative meta-study. In *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017)*, pages 299–304, 2017.

[Holzinger *et al.*, 2016] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. *Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to Solve the Traveling Salesman Problem with the Human-in-the-Loop Approach*, pages 81–95. Springer International Publishing, Cham, 2016.

[Hu *et al.*, 2014] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, Jun 2014.

[Kapoor *et al.*, 2010] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1343–1352. ACM, 2010.

[Krause *et al.*, 2014] J. Krause, A. Perer, and E. Bertini. Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623, Dec 2014.

[Krause *et al.*, 2016] 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*, CHI '16, pages 5686–5697, New York, NY, USA, 2016. ACM.

[Kulesza *et al.*, 2015] 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. ACM, 2015.

[Lee *et al.*, 2017] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105(Supplement C):28 – 42, 2017.

[Letham *et al.*, 2015] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[Lim *et al.*, 2009] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM.

[Lipton, 2016] Zachary C. Lipton. The mythos of model interpretability. In *ICML Workshop on Human Interpretability of Machine Learning*, 2016.

[Michalski, 1983] Ryszard S. Michalski. A theory and methodology of inductive learning. *Artificial Intelligence*, 20(2):111 – 161, 1983.

[Mühlbacher *et al.*, 2014] 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, Dec 2014.

[Piltaver *et al.*, 2016] Rok Piltaver, Mitja Luštrek, Matjaž Gams, and Sanda Martinčić-Ipšić. What makes classification trees comprehensible? *Expert Systems with Applications*, 62:333 – 346, 2016.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[Schreck *et al.*, 2009] Tobias Schreck, Jürgen Bernard, Tatiana von Landesberger, and Jörn Kohlhammer. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29, 2009.

[Stumpf *et al.*, 2009] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639 – 662, 2009.

[Zhou, 2005] Zhi-Hua Zhou. Comprehensibility of data mining algorithms. *Encyclopedia of Data Warehousing and Mining*, pages 190–195, 2005.