Machine Learning – the results are not the only thing that matters! What about security, explainability and fairness?

Michał Choraś[1,2,3], Marek Pawlicki[1,2], Damian Puchalski[1], and Rafał Kozik[1,2]

[1] ITTI Sp. z o.o., Poznań
[2] UTP University of Science and Technology, Bydgoszcz, Poland
[3] FernUniversitat in Hagen (FUH), Germany
chorasm@utp.edu.pl

Abstract. Recent advances in machine learning (ML) and the surge in computational power have opened the way to the proliferation of ML and Artificial Intelligence (AI) in many domains and applications. Still, apart from achieving good accuracy and results, there are many challenges that need to be discussed in order to effectively apply ML algorithms in critical applications for the good of societies. The aspects that can hinder practical and trustful ML and AI are: lack of security of ML algorithms as well as lack of fairness and explainability. In this paper we discuss those aspects and provide current state of the art analysis of the relevant works in the mentioned domains.

Keywords: Machine Learning (ML) · AI · Secure ML · Explainable ML · Fairness

1 Introduction

Recent advances in machine learning (ML) and the surge in computational power have opened the way to the proliferation of Artificial Intelligence (AI) in many domains and applications.

Still, many of the ML algorithms offered by researchers, scientists and R&D departments focus only on the the numerical quality of results, high efficiency and low error rates (such as low false positives or low false negatives). But even when such goals are met, those solutions cannot (or should not) be realistically implemented in many domains, especially in critical fields or in the aspects of life that can impact whole societies, without other crucial criteria and requirements, namely: security, explainability and fairness. Moreover, frequently the outstanding results are achieved on data that is well-prepared, crafted in laboratory conditions, and are only achievable when implemented in laboratory environments.

However, when large scale applications of AI became reality, the realization came that the security of machine learning requires immediate attention. Malicious users, called 'Adversaries' in the AI world, can skilfully influence the inputs

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_46

fed to the AI algorithms in a way that changes the classification or regression results. Regardless of the ubiquity of machine learning, the awareness of the security threats and ML's susceptibility to adversarial attacks used to be fairly uncommon and the subject has received significant attention only recently.

Apart from security, another aspect that requires attention is the explainability of ML and ML-based decision systems. Many researchers and systems architects are now using deep-learning capabilities (and other black-box ML methods) to solve detection or prediction tasks. However, in most cases, the results are provided by algorithms without any justification. Some solutions are offered as if it was magic and the Truth provider, while for decision-makers in a realistic setting the question why (the system arrives at certain answers) is crucial and has to be answered.

Therefore, in this paper an overview of aspects and recent works on security, explainability, and fairness of AI/ML systems is presented, the depiction of those concerns can be found in Fig. 1. The major contributions of the paper are: current analysis of challenges in machine learning (other than only having good numeric results) as well as state of the art analysis of works in secure ML, explainable ML and fairness.

The paper is structured as follows: in section 2 security of machine learning is discussed and an overview of recent works is provided. Several types of adversarial attacks are mentioned, such as evasion attacks, data poisoning, exploratory attacks (an example of deep learning use for exploratory attacks can be found in [1]) etc. In Section 3 a survey of fairness in ML is presented, while in Section 4 the focus is on related works in explainable machine learning. Conclusions are given thereafter.

2 Security and Adversarial Machine Learning in disinformation

Recently it has come to attention that skilfully crafted inputs can affect artificial intelligence algorithms to sway the classification results in the fashion tailored to the adversary needs [2]. This new disturbance in the proliferation of Machine Learning has been a subject riveting attention of the researches very recently, and at the time of writing this paper a variety of vulnerabilities have been uncovered [2].

With the recent spike of interest in the field of securing ML algorithms, a myriad of different attack and defence methods have been discovered; no truly safe system has been developed however, and no genuinely field-proven solutions exist [3].

The solutions known at this point seem to work for certain kinds of attacks, but do not assure safety against all kinds of adversarial attacks. In certain situations, implementing those solutions could lead to the deterioration of ML performance [2].

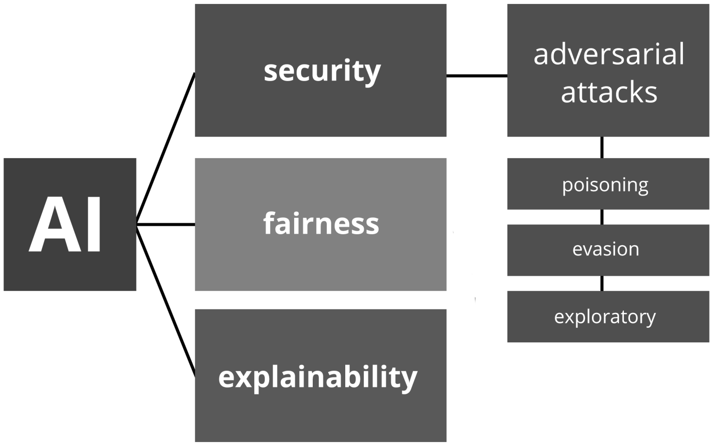

Fig. 1. Machine Learning concerns

The adversaries behaviour is affected by the extent of the knowledge the agent possesses of the target algorithm's architecture. In literature, this level of acquaintance is categorised as black box and white box [4].

While white box attacks presuppose full knowledge of the attacked algorithm; black box strikes are performed with no preceding knowledge of the model [4].

- Targeted Poisoning
- transferable clean-label attack (convex polytope attack)
- Feature Collision Attack
- one-shot kill Attack
- single poison instance attack
- watermarking
- multiple poison instance attack
- Non-Targeted Poisoning

There are a couple of known poisoning attacks featured in the literature. In [5] a method utilising the intrinsic properties of Support Vector Machines is introduced. The overarching idea is that an adversary can craft a data point that significantly deteriorates the performance of the classifier. The formulation of that data point can be, as demonstrated by the authors, defined as the solution of an optimisation problem with regard to a performance measure. The authors note that the challenge of finding the worst possible mix of label flips is not a straightforward one. The classes chosen for the flips are the ones classified with a high confidence, this should result in a significant impact on SVM accuracy[5].

The authors of [6] express in their paper an illustration of how an attacker could manipulate ML in spam filtering by meddling with the data to either subject the user to an ad, or stop the user from receiving genuine communication. The efficiency of those attacks is illustrated.

DOI: 10.1007/978-3-030-50423-6_46

The authors use an algorithm called SpamBayes for their research. SpamBayes takes the head and body of a message, tokenizes them and scores the spam to classify it as spam, ham or unsure. With this established, the paper presents a dictionary attack, in which the algorithm is subjected to an array of spam e-mails containing a set of words that are likely to be present in genuine communication. When those are marked as spam, the algorithm will be more likely to flag legitimate mail as spam. This particular attack comes in two variations: a procedure where the attack mail simply contains the whole dictionary of the English language, called the 'basic dictionary attack' and a more refined approach, where the attack is performed with the use of a message containing word distribution more alike the users message distribution, along with the colloquialisms, misspellings etc. In this particular case the authors propose a pool of Usenet newsgroup posts. The other evaluated attack is geared towards blocking a specific kind of e-mail - a causative targeted availability attack, or a focused attack. In this scenario the adversary spam the user with messages containing words that are likely to appear in a specific message. With SpamBayes retrained on these messages it is then predisposed to filtering a distinct, genuine communication as spam. This could eliminate a competing bid of a rival company, for example. Including the name of a rival company in spam e-mails, their products or the names of their employees could achieve that objective. The authors indicate that using the dictionary attacks can neglect the feasibility of a spam filter with only 1% of retraining dataset controlled, and a masterfully crafted focused attack can put a specific message in the spam box 90% of the time.

In [7] a framework for the evaluation of security of feature selection algorithms is proposed. The framework follows the outline defined by [8] in which the authors evaluate the attacker's goal, the extent of the adversary's knowledge of the workings of the algorithm, and their capability in data manipulation. The goal of the malicious user is either targeted or indiscriminate, and it aims to infringe on one or more of the well-known infosec triad items: availability, integrity or privacy. The specific acquaintance of the adversary with the workings of the system can be one of the following:

- knowledge of the training data (partial or full)
- knowledge of the feature representation (partial or full)
- knowledge of the feature selection algorithm
- Perfect Knowledge (worst case scenario)
- Limited Knowledge

With regard to the attacker's capability, in the case of causative (poisoning) attacks, the attacker can usually influence just a subsection of the training set. The adversary has to bear in mind that the labeling process varies in different use cases, with the use of honeypots and anti-virus software giving setting the constraints of the datapoints that have to be crafted in the malware detection example.

The authors evaluate the robustness of 3 widely-used feature selection algorithms: Lasso, Ridge Regression and the Elastic Net against poisoning attacks with regard to the percentage of injected poisoned data points. The results show that poisoning 20% of the data inflates the classification error 10-fold. In addition to the influence on classification, the authors notice that even with a minute amount of poisoning samples the stability index drops to zero. This means that the attacker can influence feature selection.

In [9] the authors investigate a poisoning attack geared towards targeting specific test instances with the ability to fool a labelling authority, which they name 'clean-label' attacks. Their work does not assume knowledge of the training data, but does require the knowledge of the model. It is an optimisation-based attack for both the transfer-learning and end-to-end DNN training cases. The overall procedure of the attacks, called 'Poison Frogs' by the authors, is as follows: the basic version of this attack starts with choosing the target datapoint, then making alterations to that datapoint to make it seem like it belongs to the base class. A poison crafted that way is then inserted into the dataset. The objective is met if the target datapoint is classified as the base class at test time. Arriving at a poisonous datapoint to be inserted into the training set comes as a result of a process called 'feature collision'. It is a process that exploits the nonlinear complexity of the function propagating the input through the second-to-last layer of the neural network to find a datapoint which 'collides' with the target datapoint, but is also close to the base class in the feature space. This allows the poisoned datapoint to bypass the scrutiny of any labelling authority, and also remain in the target class distribution. The optimisation is performed with a forward-backward-splitting iterative procedure.

The [10] paper evaluates a poisoning procedure geared towards poisoning multi-class gradient-decent based classifiers. To this end the authors utilize the recently proposed back-gradient optimization. This approach allows for a replacement of one of the optimisation problems with a set of iterations of updating the parameters.

The authors introduce an attack procedure to poison deep neural networks taking into consideration the weight updates, rather than training a surrogate model trained on deep feature representations. They demonstrate the method on a convolutional neural network (CNN) trained on the well-known MNIST digit dataset, a task which requires the optimisation of over 450000 parameters. They find that deep networks seem more resilient to poisoning attacks than regular ML algorithms, at least in conditions of poisoning under 1% of the data. The authors also conduct a transferability experiment in which they conclude that poisons crafted against linear regression (LR) algorithm are ineffective against a CNN, and poisons crafted against a CNN have a similar effect on LR as random label flips. A more comprehensive assessment of the effects of poisoning attacks crafted against deep neural networks with the use of the back-gradient algorithm is necessary. The notion of transferability of adversarial attacks is explored in depth in [11], where the authors conclude that, for evasion attacks in their case, many attacks could be effectively used across models trained with different techniques, and prove their findings by attacking classifiers hosted by Amazon and Google without any knowledge of the attacked models. An in-depth analysis of the transferability of both evasion and poisoning attacks is performed in [12].

DOI: 10.1007/978-3-030-50423-6_46

[13] investigates poisoning attacks carried out by an attacker with full knowledge of the algorithm. The assumption is that the adversary aims to poison the model with the minimum amount of poisoning examples. The attacker function is defined as a bilevel optimisation problem. The authors notice that this function is similar to machine teaching, where the objective is to have maximum possible influence over the subject by carefully crafting the training dataset. The authors point to the mapping of a teacher to the attacker and from a student to the AI algorithm. The paper thus offers economical solutions to the bilevel optimisation problem present in both fields. Essentially, the authors suggest that, under certain regulatory conditions, the problem can be reduced with the use of Karush-Kuhn-Tucker theorem (KKT) to a single-level constrained optimization problem. Thus, a formal framework for optimal attacks is introduced, which is then applied in 3 different cases - SVM, Linear Regression and logistic regression. In [14] the authors propose a way of bypassing the gradient calculation by partially utilising the concept of a Generative Adversarial Network (GAN). In this approach an autoencoder is applied to craft the poisoned datapoints, with the loss function deciding the rewards. The data is fed to a neural network, and the gradients are sent back to the generator. The effectiveness of their method is tested thoroughly on the well-known MNIST and CIFAR-10 datasets. The chosen architecture is a two-layer feed forward neural network with recognition accuracy of 96.82% on the MNIST dataset, and for CIFAR-10 a convolutional neural network with two convolutional layers and two fully-connected layers, with the accuracy of 71.2%. For demonstrative purposes, one poisoned datapoint is injected at a time. The authors conclude that the generative attack method shows improvement over the direct gradient methods and stipulate that it is viable for attacking deep learning and its datasets, although more research is required. A targeted backdoor attack is proposed in [15]. The premise of the method is to create a backdoor to an authentication system based on artificial intelligence, allowing the adversary to pass the authentication process by deceiving it. The poisoning datapoints are created specifically to force an algorithm to classify a specific instance as a label of the attacker's choice. The authors propose a method that works with relatively small poison samples and with the adversary possessing no knowledge of the algorithm utilised. This claim is backed up by a demonstration of how inserting just 50 samples gets a 90% success rate.

3 Fairness in Machine Learning

Fairness in Machine Learning (or Artificial Intelligence in a broader sense) is a concept which is getting an increased amount of attention with the growing popularity of AI in different society-impacting applications. Fairness in AI is mainly ethically and legally motivated. It is also a fertile ground to spread politically motivated disinformation, when used maliciously.

The background of the fairness concept in AI results from the misinformed, but widespread perception of Artificial Intelligence and AI/ML-based decision making as fully objective. In practice, the fairness of AI-driven decisions depends

DOI: 10.1007/978-3-030-50423-6_46

highly on the data provided as the input to learning algorithms. This data can be (and often is) biased due to several reasons: 1) bias of human operators providing this data as input, resulting e.g. in biased labeling of samples, 2) data unbalance/misrepresentation of e.g. specific minority groups, 3) historical bias (discrimination) [16].

In addition, in [17] there is a list of potential causes of bias in training datasets leading to unfairness in AI:

- Skewed sample – misrepresentation of training data in some areas that evolves over time. In that way the future observations confirm biased prediction and misrepresented data samples give less chances for contradicting observations.
- Tainted examples - the bias existing in the old data caused by human bias is replicated by the system trained on this data.
- Limited features - reliability of some labels from a minority group (e.g. unreliably collected or less informative) impacts the system and may cause the lower accuracy for the predictions related the minority group.
- Sample size disparity – causing difficulties in building a reliable model of the group described by an insufficient data sample.
- Proxies – correlation of sensitive biased attributes (even if not used to train an ML system, encrypted, etc.) with other features preserves a bias in predicted output.

All those reasons result in inaccurate decisions based on (or related to) sensitive attributes such as gender, race or others. A decision-making process is affected by disparate treatment if its decisions are based on these sensitive attributes. In summary, the definition of unfair machine learning process can be formulated as a situation in which an output tends to be disproportionately benefitting (or unfair) towards the group characterized by certain sensitive attribute values. However, this commonly used definition is too abstract to reach a consensus on the mathematical formulations of fairness. The majority of definitions of fairness in ML include the following elements [16][18][19]: group fairness (including demographic parity, equalized odds and predictive rate parity), unawareness, individual fairness and counterfactual fairness.

The author of [16] discusses some solutions to address each of these elements:

- Not including sensitive attributes' values in a training dataset (addressing the unawareness and disparate treatment of subjects). The challenge here is in the existence of the proxies, i.e. non-sensitive attributes used to train ML system highly correlated with eliminated sensitive attributes.
- Statistical parity of different groups in the training sample – e.g. application of the so-called 80% rule - the size of the sample belonging to the group with the lowest selection rate should be at least 80% in comparison to the mostly represented group (proportion should be higher than 4/5). This could prevent from extreme misrepresentation of minor groups.
- Optimal adjustment of learned predictor to reduce discrimination against a specified sensitive attribute in supervised learning according to the equalized odds definition [20].

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_46

- Replacing the original value of the sensitive attribute by the counterfactual value propagated “downstream” in the causal graph. This addresses counterfactual fairness and provides a way to check and explain the possible impact of bias via a causal graph [21]. As pointed in [16], in practice in many applications it is hard to build a causal graph and the elimination of correlated attributes can result in significantly decreased accuracy of prediction.

State of the art on algorithm for fair ML

As the fairness in ML/AI is a trending topic, there are many algorithms focusing on improving fairness in ML described in the literature. Most of them fall into three categories: preprocessing, optimization at training time, and postprocessing [16]. In general, algorithms belonging to the same category are characterized by common advantages and flaws.

Preprocessing The idea is based on building a new representation of input data by removing the information correlated to the sensitive attribute and at the same time preserving the remaining input information as much as possible. The downstream task (e.g. classification, regression, ranking) can thus use the “cleaned” data representation and produce results that preserve demographic parity and individual fairness. In [22] authors use the optimal transport theory to remove disparate impact of input data. They also provide numerical analysis of the database fair correction. In [23] authors propose a learning algorithm for fair classification addressing both group fairness and individual fairness by obfuscating information about membership in the protected group. Authors of [24] propose a model based on a variational autoencoding architecture with priors that encourage independence between sensitive and latent factors of data variation. To remove any remaining dependencies an additional penalty term based on the “Maximum Mean Discrepancy” (MMD) measure is additionally introduced. A statistical framework for removing information about a protected variable from a dataset is presented in [25], along with the practical application to a real-life dataset of recidivism, proving successful predictions independent of the protected variable, with the predictive accuracy preserved. [26] proposes a convex optimization for learning a data transformation with three goals: controlling discrimination, limiting distortion in individual data samples, and preserving utility.

The authors of [27] and [28] evaluate the use of dataset balancing methods for data augmentation to counter fairness issues. Both papers report positive results.

Optimization at training time Data processing at training time provides good performance on accuracy and fairness measure and ensures higher flexibility in optimizing the trade-off between these factors. The author of [16] describes that the common idea that can be found in the state-of-the-art works falling into this category of algorithms is to add a constraint or a regularization term to the existing optimization objective. Recent works considering algorithms to ensure ML fairness applied at the training time include: [29] where the problem of learning a non-discriminatory predictor from a finite training set is studied to preserve “equalized odds” fairness, [30] and [21] where a flexible mechanism to design fair classifiers by leveraging a novel intuitive measure of the decision boundary (un)fairness is introduced, and [31] that addresses the problem of reducing the fair classification to a sequence of cost-sensitive classification problems, whose solutions provide a randomized classifier with the lowest (empirical) error subject to the desired constraints.

The disadvantages of abovementioned approaches include the fact that these methods are highly task-specific and they require a modification of the classifier, which can be problematic in most applications/cases. **Post-processing** The post-processing algorithms are focused on editing the posteriors to satisfy the fairness constraints and can be applied to optimize most of fairness definitions except the counterfactual fairness. The basic idea is to find a proper threshold using the original score function for each group. An exemplary recent work that falls into this category is the publication [20], in which the authors show how to optimally adjust any learned predictor to remove the discrimination according to the "equal opportunity" definition of fairness, with the assumption that data about the predictor, target, and membership in the protected group are available.

The advantage of post-processing mechanisms is that retraining/changes are not needed for the classifier (the algorithm can be applied after any classifier)[16]. **Summary** The author of [31] provides an experimental comparison of the selected algorithms applied to reduce unfairness using four real-life datasets with one or two protected sensitive attributes (gender or/and race). The selected methods include preprocessing, optimization at training time and post-processing approaches. The methods that achieve the best trade-off between accuracy and fairness are those falling into the optimization at training time category, while the advantage related to the implementation of preprocessing and post-processing methods is the preservation of fairness without modifying the classifiers. In general, experimental results prove the ability to significantly reduce or remove the disparity, in general not impacting the classifier's accuracy for all the methods [16][31].

4 AI explainability and interpretability

The aspects of explainability and interpretability are trending topics in the area of Machine Learning and Artificial Intelligence in general as well. As discussed in [32] and [33] these two terms – explainability and interpretability tend to be used (also in literature) interchangeably, however despite the fact that they are related concepts, there are some minor differences in their meanings. Interpretability addresses the aspects related to observation of AI system outputs. Interpretability of AI system is higher, if the changes of the systems outputs in result of changing algorithmic parameters are more predictable. In other words, system interpretability is related to the extent to which a human can predict the results of AI systems based on different inputs. On the contrary, explainability is related to the extent to which a human can understand and explain (literally) the internal mechanics of an AI/machine learning system. In its simplest form, the definition of explainability refers to an attempt to provide insights

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_46

into the predictor's behavior [34]. According to [33], nowadays, attempts to define these concepts are not enough to form a common and monolithic definition of explainability and interpretability and to enable their formalization. It is also worth mentioning, that the "right to explanation" in the context of AI systems directly affecting individuals by their decisions, especially legally and financially is one of the subjects of the GDPR [35].

Different scientific and literary sources focus on surveying and categorisation of methods and techniques addressing explainability and interpretability of decisions resulting from AI systems use. [32] discusses the most common practical approaches, techniques and methods used to improve ML interpretability and enable more explainable AI. They include, among others, algorithmic generalization, i.e. shifting attitude from case-specific models to more general ones. Another approach is paying attention to feature importance, described also in [34] as the most popular technique addressing ML explainability, also known as feature-level interpretations, feature attributions or saliency maps. Some of feature importance-based methods found in the literature are perturbation-based methods based on Shapley values adapted from the cooperative game theory. In the explainability case, Shapley values are used to attain fair distribution of gains between players, where a cooperative game is defined between the features. In addition, some recent works [32][36] show that adversarially trained models can be characterised by increased robustness but also provide clearer feature importance scores, contributing to improved prediction explainability. Similar to the feature importance way, counterfactual explanations [34] is a technique applied in the financial and healthcare domains. Explanations using this method are based on providing point(s) and values that are close to the input values for which the decision of the classifier possibly changes (case-specific threshold values). Another method used for increasing explainability of AI-based predictions is LIME (Local Interpretable Model-Agnostic Explanations) based on approximation of the model by testing it, then applying changes to the model and analysis of the output. DeepLIFT (Deep Learning Important Features) model is used for the challenge-based analysis of deep learning/neural networks. As described in [32] DeepLift method is based on backpropagation, i.e. digging back into the feature selection inside the algorithm and "reading" neurons at subsequent layers of network. The authors of [37] evaluate the use of influence functions as a way of selecting specific training samples responsible for a given prediction, a paradigm known as prototype-based explanation.

In the literature one can find different attempts of categorization of the methods aimed at increased explainability of AI. Integrated/Intrinsic and post-hoc explainability methods [33] [38] is one of such categorization. Intrinsic explainability in its simplest form is applicable to some basic variants of the low complexity models, where the explanation of a simple model is the model itself. On the other hand, more complex models are explainable in a post-hoc way, providing explanations after the decision and using techniques such as feature importance, layer-wise relevance propagation, or the mentioned Shapley values.

Similar categorization is given in [38] where in-model (integrated/intrinsic) and post-model (post-hoc) methods exist alongside additional pre-model interpretability methods. Pre-model methods are applicable before building (or selection) of the ML model and are strictly related to the input data interpretability. They use mainly classic descriptive statistical methods, such as PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and clustering methods such as k-means. Another criterion described in [38] is the differentiation into model-specific and model-agnostic explanation methods. In the majority of cases model-specific explanation methods are applicable to the intrinsically interpretable models (for example analysis and interpretation of weights in a linear model), while model-agnostic methods can be applied after the model and include all post-hoc methods relying on the analysis of pairs of feature input and output. Alternative criterion based on explanation methods is described in [39]. In such differentiation, methods are categorized based on type of explanation that the given method provides, including: feature summary (providing statistic summary for each feature with their possible visualization), model internals (for intrinsic explainable or self-explainable models), data point (example-based models) and a surrogate intrinsically interpretable model - that is trained to approximate the predictions of a black box model.

According to [38] and [40] explanation models can be evaluated and compared using qualitative and quantitative metrics, as well as by comparison of the explanation method's properties, including its expressive power, translucency (model-specific vs. model-agnostic), portability (range of applications) and computational complexity. On the other hand, individual explanations can be characterized by accuracy, fidelity, consistency (similarity of explanations provided by different models), stability, comprehensibility, certainty, to list most relevant ones. According to the literature we can also distinguish qualitative and quantitative indicators to assess the explanation models. Factors related to quality of explainability are: form of the explanation, number of the basic units of explanation that it contains, compositionality (organization and structure of the explanation), interactions between the basic explanation units (i.e. intuitiveness of relation between them), uncertainty and stochasticity. Quantitative indicators are presented in some works (e.g. [38][41][42]). The most common metrics used to quantify the interpretation of ML models are identity, separability and stability. These three factors provide the information on to what extent identical, non-identical and similar instances of predictions are explained in identical, non-identical and similar way, respectively. In addition, according to [41] the explanation should be characterized by high completeness (coverage of the explanation), correctness and compactness. However, these indicators are applicable only to simple models (rule-based, example-based).

5 Conclusions

In this paper recent research in secure, explainable and fair machine learning was surveyed. The high number of related works shows that those aspects are

DOI: 10.1007/978-3-030-50423-6_46

becoming crucial. At the same time an increasing number of researchers are aware that in machine learning the numeric results are not the only thing that matters. This work is a part of the SAFAIR Programme (Secure and Fair AI Systems for Citizens) of the H2020 SPARTA project that focuses on security, explainability, and fairness of AI/ML systems, especially in the cybersecurity domain. Moreover, the same aspects (secure, fair and explainable ML) are a part of the project SIMARGL focusing on detection on malware by advanced ML techniques. We believe that even more projects will contain the work on secure and explainable machine learning, and that this survey will be helpful and might inspire more researchers in ML community to seriously consider those aspects.

6 Acknowledgement

This work is funded under the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892. This work is partially funded under SIMARGL project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 833042.

References

1. Michał Choraś, Marek Pawlicki, and Rafał Kozik. The feasibility of deep learning use for adversarial model extraction in the cybersecurity domain. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 353–360. Springer, 2019.
2. Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.
3. Xiaofeng Liao, Liping Ding, and Yongji Wang. Secure machine learning, a brief overview. In *2011 Fifth International Conference on Secure Software Integration and Reliability Improvement-Companion*, pages 26–29. IEEE, 2011.
4. Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.
5. Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.
6. Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles A Sutton, J Doug Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. *LEET*, 8:1–9, 2008.
7. Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, pages 1689–1698, 2015.
8. Battista Biggio, Giorgio Fumera, and Fabio Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(07):1460002, 2014.

DOI: 10.1007/978-3-030-50423-6_46

9. Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018.
10. Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec 17*. ACM Press, 2017.
11. 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*, 2016.
12. Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 321–338, 2019.
13. Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
14. Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. Generative poisoning attack method against neural networks. *arXiv preprint arXiv:1703.01340*, 2017.
15. 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.
16. Ziyuan Zhong. A tutorial on fairness in machine learning, Jul 2019.
17. Solon Barocas and Andrew D. Selbst. Big datas disparate impact. *SSRN Electronic Journal*, 2016.
18. Pratik Gajane and Mykola Pechenizkiy. On Formalizing Fairness in Prediction with Machine Learning. *arXiv e-prints*, page arXiv:1710.03184, Oct 2017.
19. Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness - FairWare 18*. ACM Press, 2018.
20. Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
21. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. In *Proceedings of the 26th International Conference on World Wide Web - WWW 17*. ACM Press, 2017.
22. Eustasio Del Barrio, Fabrice Gamboa, Paula Gordaliza, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. *arXiv preprint arXiv:1806.03195*, 2018.
23. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.
24. Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.
25. Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.
26. Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_46

27. Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, page 24, 2018.
28. Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.
29. Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.
30. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.
31. Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.
32. R Gall. Machine learning explainability vs interpretability: Two concepts that could help restore trust in ai. *KDnuggets News*, 19(1), 2019.
33. Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, may 2018.
34. Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. *arXiv preprint arXiv:1909.06342*, 2019.
35. Albert C. We are ready for machine learning explainability?, Jun 2019. url: https://towardsdatascience.com/we-are-ready-to-ml-explainability-2e7960cb950d, last visited: 2020-31-03.
36. Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.
37. Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.
38. Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, jul 2019.
39. Christoph Molnar. A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book/(last visited: 28.03. 2019)*, 2018.
40. Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and Machine Learning*, pages 159–175. Springer International Publishing, 2018.
41. Wilson Silva, Kelwin Fernandes, Maria J. Cardoso, and Jaime S. Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140. Springer International Publishing, 2018.
42. Milo Honegger. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*, 2018.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_46