RITUAL: a Platform Quantifying the Trustworthiness of Supervised Machine Learning

Alberto Huertas Celdrán[1], Jan Bauer[1], Melike Demirci[1], Joel Leupp[1], Muriel Figueredo Franco[1], Pedro M. Sánchez Sánchez[2], Gérôme Bovet[3], Gregorio Martínez Pérez[2], Burkhard Stiller[1]
[1]Communication Systems Group CSG, University of Zürich UZH, CH–8050 Zürich, Switzerland
[huertas, bauer, demirci, leupp, franco, stiller]@ifi.uzh.ch
[2]Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain
[pedromiguel.sanchez, gregorio]@um.es
[3]Cyber-Defence Campus, armasuisse Science & Technology, CH–3602 Thun, Switzerland
gerome.bovet@armasuisse.ch

***Abstract*—This demo presents RITUAL, a platform composed of a novel algorithm and a Web application quantifying the trustworthiness level of supervised Machine and Deep Learning (ML/DL) models according to their fairness, explainability, robustness, and accountability. The algorithm is deployed on a Web application to allow users to quantify and compare the trustworthiness of their ML/DL models. Finally, a scenario with ML/DL models classifying network cyberattacks demonstrates the platform applicability.**

***Index Terms*—Artificial Intelligence, Trust, Fairness, Explainability**

I. Introduction

Over the last decade, Artificial Intelligence (AI) improved vital challenges of communication networks. Some examples are the detection of cyberattacks, the real-time reconfiguration of network services, or the optimization of traffic routes [1]. In the soon future, the AI and networking alliance will be even more effective with AI integrated into the core of the 6G mobile network protocol stack [2]. Thus, it is crucial to have mechanisms to quantify the trustworthiness level of AI systems and their predictions.

Recently, the research community has agreed on the importance of fairness, explainability, robustness, and accountability as pillars to trust AI systems [3]. Concerning fairness, bias is one of the main issues against trusting AI systems. Machine and Deep Learning (ML/DL) models could be unfair due to *i)* biased training data, *ii)* unbalanced or lack of training data, or *iii)* discrimination of *protected groups*, among others [4]. Explainability is another pillar for trusting AI systems that consists of understanding how ML/DL models come to their conclusions. The algorithm class, features importance, or model complexity are some of the key aspects to explain model predictions [5]. Robustness refers to the model ability to deal with adversarial attacks. Therefore, to ensure trusted ML/DL models, their predictions must be stable and robust, even when adversaries are present [6]. Finally, the training methodology, and its accountability in terms of data splitting, pre-processing, or normalization are key aspects providing valuable insights to trust AI systems.

Trusted AI is an emerging field that needs more effort despite the contributions of work related to the previous four pillars. In particular, the literature lacks a comprehensive and unified collection of relevant metrics able to quantify trusted ML/DL models. Furthermore, existing solutions are isolated and only focus on detecting or mitigating individual issues per pillar. Therefore, there is no solution combining metrics of different pillars relevant for trusted AI and computing a global trustworthiness level of ML/DL models.

This demo paper presents *RITUAL (platfoRm quantIfying Trustworthiness in sUpervised mAchine Learning)*, which is composed of a Web application and an extensible and parameterized algorithm to quantify the trustworthiness level of supervised ML/DL models according to a set of relevant metrics dealing with the fairness, robustness, explainability, and accountability pillars [7]. RITUAL allows users to compute and compare the trustworthiness level of their ML/DL models using a user-friendly Web-based interface. The suitability of the platform is demonstrated in a scenario focused on detecting cyberattacks on Internet of Things (IoT) devices.

II. RITUAL Platform

The RITUAL platform is composed of an algorithm quantifying the trustworthiness of ML/DL models and a Web application. The pillars, metrics, and life-cycle of the algorithm are shown in Figure 1. After reviewing the literature, the following metrics per pillar haven been selected as relevant for trusted AI.

A. Fairness Metrics [8]

- *Underfitting*: calculates the difference between train and baseline performance.
- *Overfitting*: computes the difference between train and test performance, giving the model generalization.
- *Class Balance*: measures the ratio of samples belonging to different classes in the training dataset.
- *Statistical Parity Difference*: computes the spread between the percentage of samples receiving a favorable outcome for protected and unprotected samples groups.
- *Equal Opportunity Difference*: measures the spread between true positive rate (TPR) and false positive rate (FPR) between different groups.

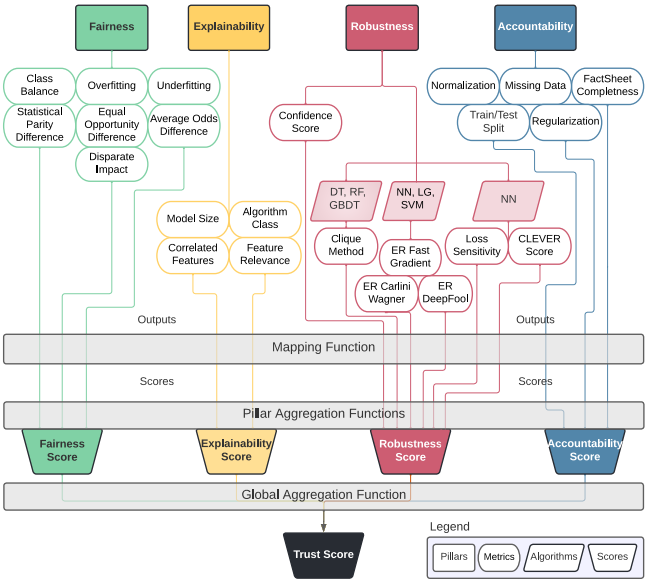

Figure 1: Overview of the RITUAL Algorithm

- *Average Odds Difference*: calculates the mean absolute difference in TPR and FPR between protected and unprotected groups.
- *Disparate Impact*: measures the ratio of a protected and unprotected groups receiving a favorable prediction.

B. Explainability Metrics [9]

- *Algorithm Class*: calculates the model explainability degree according to the algorithm type and its complexity.
- *Correlated Features*: measures the percentage of highly correlated features.
- *Feature Relevance*: computes the percentage of irrelevant features for a set of predictions.
- *Model Size*: calculates the number of parameters used by models.

C. Robustness Metrics [10]

- *Confidence Score*: measures the probability of predicting correctly a given sample.
- *Loss Sensitivity*: calculates the largest variation of a Neural Network output under a small change in its input.
- *Cross Lipschitz Extreme Value for Network Robustness (CLEVER) Score*: measures the minimal perturbation that is needed to change a classification outcome.
- *Clique Method*: finds the exact minimal adversarial perturbation or a guaranteed lower bound of it.
- *Empirical Robustness (ER)*: measures the average minimal perturbation that needs to be introduced to change the model prediction.

D. Accountability Metrics [11]

- *Train/Test Split*: measures the ratio between the number of samples used for training and testing.
- *Missing Data*: evaluates how missing values of features of the training dataset are handled.
- *Normalization*: evaluates if some models have been trained with normalized or non-normalized data.
- *Regularization*: measures if the ML/DL model used generalization techniques during training.
- *FactSheet Completeness*: measures if the FactSheet includes all necessary information that stakeholders need in order to trust the model and its predictions.

Each metric receives as input the *i)* training and testing datasets, *ii)* trained ML/DL model, and *iii)* metadata of the training methodology (called factsheet [11]). Then, the algorithm evaluates if the inputs fulfill the conditions of each metric. If so, each metric is independently computed according to its formula and input data [12].

The metrics outputs cannot be interpreted as trust scores because they have different data types, scales, and meanings. Therefore, each metric output must be interpreted and translated into a standard trust score using a mapping function. The proposed trust score for all metrics ranges from one to five, where one corresponds to the worst score, and five represents the best score. The mappings from metrics outputs to trust scores are predefined by the RITUAL algorithm according to good practices indicated in the literature. However, to avoid arbitrary decisions adding biases, the mapping function is parameterized and can be fine-tuned by stakeholders according to the data domain, metric, or scenario.

The next step consists of aggregating all the metrics scores of each pillar and calculating a score per pillar. The RITUAL algorithm proposes a weighted approach where each metric has particular importance in the pillar score. It is up to discuss whether all metrics are equally important and how weighted they should be. Because of that, default weights for every metric are defined, but stakeholders can modify them according to the scenario characteristics.

Then, the four pillar scores are aggregated into a global trust score, which is the return value of the RITUAL algorithm. Computing the global trust score is done analog to calculating the pillars scores. Independent weights are assigned to each pillar, and the global trust score is the weighted average of each pillar. Since the importance of each pillar depends on the scenario, the predefined configuration of the algorithm (equal importance per pillar) can be modified by stakeholders.

Finally, to allow users to apply the algorithm, the RITUAL platform provides a Web application. This application enables stakeholders to *i)* upload supervised ML/DL models, datasets, and factsheet, *ii)* calculate and graphically see the trustworthiness level of the uploaded models, *iii)* fine-tune the algorithm parameters according to the scenario requirements, and *iv)* compare the trustworthiness levels of ML/DL models.

III. Demonstration

A scenario focused on classifying different network data leakage attacks affecting a Raspberry Pi has been defined to demonstrate the suitability of the RITUAL platform.

First, one dataset modeling the Raspberry device behavior has been created. The dataset contains nine features belonging to the following events families: *network packets scheduled, bytes transmitted, bytes received, TCP probe events, network buffers, socket creation and destruction*. In addition, four different device behaviors (labels) are contained in the

dataset: (i) *Normal behavior*; (ii) *Behavioral data leakage*, which periodically opens an ssh connection from a command and control (C&C) server to the device and leak device behavioral data; (iii) *Cryptographic material leakage*, where the C&C reads the content of sensitive files of the device; and (iv) *Sensitive application data leakage*, where the device leaks different amounts of sensitive data. In total, the dataset contains 3972 samples. After that, a common ML/DL training pipeline, with 90/10% data splitting strategy and data min-max normalization, is performed to train a Support Vector Machine (SVM) and a k-Nearest Neighbor (K-NN) model able to classify the cyberattacks.

Table I: RITUAL Platform Output

Model	Accuracy	Trust Score	Fair.	Exp.	Rob.	Account.
SVM	0.97	**3.2**	3.2	2.4	4.0	3.4
K-NN	0.97	**3.4**	3.2	3.0	4.0	3.2

As can be seen in Table I, both SVM and k-NN achieved 97% accuracy. However, after uploading the previous models, dataset, and factsheet to the RITUAL platform, SVM achieved a 3.2 (of 5) trust score, while k-NN obtained 3.4 due to a higher explainability score (see Table I). Figure 2 shows the performance metrics, properties, overall trustworthiness score, and pillars scores of the SVM model. Besides, Figure 3 shows the difference between both solutions in terms of explainability.

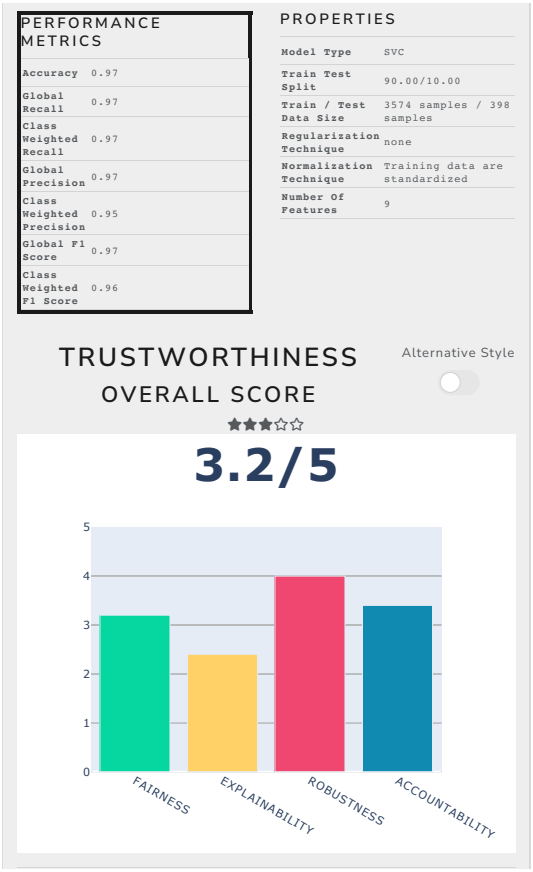

Figure 2: Overall Trustworthiness Score for SVM

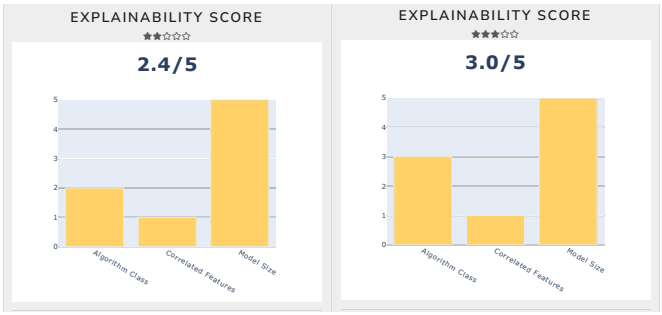

Figure 3: Explainability Metrics (left: SVM, right: kNN)

References

[1] A. Kaloxylos, A. Gavras, D. Camps Mur, M. Ghoraishi, and H. Hrasnica. AI and ML—Enablers for Beyond 5G Networks. *Zenodo, Honolulu, HI, USA, Technical Report*, 2020.

[2] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu. Artificial-Intelligence-Enabled Intelligent 6G Networks. *IEEE Network*, 34(6):272–280, 2020.

[3] J. Wing. Trustworthy AI. *Communications of the ACM*, 64(10):64–71, September 2021.

[4] G. Saposnik, D. Redelmeier, C. C. Ruff, and P.N. Tobler. Cognitive Biases Associated with Medical Decisions: A Systematic Review. *BMC Medical Informatics and Decision Making*, 16(1):1–14, 2016.

[5] S. Dash, O. Günlük, and D. Wei. Boolean Decision Rules via Column Generation. In *International Conference on Neural Information Processing Systems (NeurIPS 2018)*, page 4660–4670, Montréal, Canada, 2018.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] A. Huertas Celdrán, J. Bauer, M. Demirci, J. Leupp, M. F. Franco, B. Stiller, G. Bovet, P. M. Sánchez Sánchez, and G. Martínez Pérez. Web Application to Quantify Trusted AI, 2022. Available at https://www.trusted-ai.net, Credentials - user: trusted-ai; password: mp@csg2021.

[8] R. K. E. Bellamy et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943*, 2018.

[9] V. Arya et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[10] M.-I. Nicolae et al. Adversarial Robustness Toolbox v1.0.0. *arXiv preprint arXiv:1807.01069*, November 2019.

[11] J. Richards, D. Piorkowski, M. Hind, S. Houde, and A. Mojsilovic. A Methodology for Creating AI FactSheets. *arXiv preprint arXiv:2006.13796*, 2020.

[12] A. Huertas Celdrán, J. Bauer, M. Demirci, J. Leupp, M. F. Franco, B. Stiller, G. Bovet, P. M. Sánchez Sánchez, and G. Martínez Pérez. Trusted-AI Git Repository, 2022. Available at https://github.com/JoelLeupp/Trusted-AI.