eXplainable and Reliable AI Approaches to Trustworthy AI

Alberto Carlevaro[1,2], Marta Lenatti[1], Martina Mammarella[1], Marco Muselli[1,3], Sara Narteni[1,4], Vanessa Orani[1], Alessia Paglialonga[1], Fabrizio Dabbene[1], Maurizio Mongelli[1]

[1] National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italy;

[2] University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), 16145 Genoa, Italy;

[3] Rulex Inc., Newton, MA, USA;

[4] Politecnico di Torino, DAUIN department, 10129, Turin, Italy;

name.surname@ieiit.cnr.it

Abstract

In the last decade, the Information and Systems Engineering (ISE) group of the CNR institute IEIIT has contributed extensively to the growth of the scientific community in the field of AI. This short paper aims at summarizing our best contributions in the field of trustworthy AI, focusing on reliable approaches.

1 Introduction

1.1 Research Team

The ISE research group, based in Turin, Milan and Genoa, is part of the Institute of Electronics, Information Engineering and Telecommunications (IEIIT) of the National Research Council of Italy (CNR) and involves researchers with multidisciplinary backgrounds.

1.2 Research Themes

The focus of our research is to provide a set of tools in the area of trustworthy AI, which can be tailored to a variety of use cases including cyber-physical systems, cybersecurity problems and applications in the medical field. The development of AI based solutions is pursued in compliance with the relative EU requirements [1], with a special focus to *Transparency* and *Technical robustness and safety* requirements. Specifically, *Transparency* is guaranteed by either the use of fully explainable models (e.g., the Logic Learning Machine), or the use of non-natively explainable models coupled with both local (i.e., LIME, counterfactual explanations) and global post-hoc explainability techniques [Guidotti *et al.*, 2019]. The *Technical robustness and safety* requirements fulfilment is instead addressed by assuring a satisfactory level of accuracy, reliability and robustness to harmful attacks (i.e., adversarial attacks) of the proposed solutions.

[1] European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019, https://data.europa.eu/doi/10.2759/177365

2 Methodology

This section lists some of the most used techniques developed by our team to date. Each has been published or is under submission.

2.1 Logic Learning Machine

Logic Learning Machine (LLM) is a supervised method, developed in [Muselli, 2006], and now constituting the main core of the Rulex Analytics tools. Considering classification tasks, the LLM builds a set of intelligible rules in **if-then** format, predicting an output class value based on the logical product of conditions on input variables. Each rule can be evaluated through covering and error metrics: the first expresses the proportion of how many samples are correctly classified by the rule, while the error tells how many points are wrongly classified.

Based on covering and error, a relevance can be associated to each input feature, leading to a *feature ranking*, or to each subset of its values, hence finding the *value ranking*.

2.2 Reliability from LLM

We consider binary classification problems through LLM, where an input features vector x has to be associated with one out of two classes: in our context, we refer to a safe class, $y = 0$, and an unsafe class, $y = 1$.

In our recent work [Narteni *et al.*, 2021], we defined three methods that exploit feature and value ranking to design *safety regions* with zero FNR (i.e., unsafe class points wrongly classified as safe), saving the highest number of safe points as possible: *reliability from outside*, *reliability from inside* and *LLM with zero error*.

The first two methods share the same methodological approach, as summarized in **Algorithm 2**.

Let $\mathcal{X}$ be the input dataset, with D_1 samples for class $y = 1$ and D_0 samples for class $y = 0$.

Algorithm 2 `ReliabilityFromLLM`
Inputs: dataset $\mathcal{X}$; number of features N_{FR};
candidate perturbations $\Delta = (\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_{N_{FR}})$, $\boldsymbol{\delta}_j = (\delta_{s_j}, \delta_{t_j})$, $j = 1, ..., N_{FR}$.

1. Apply LLM on $\mathcal{X}$;
2. Select N_{FR} features from feature ranking;
3. Find $[s_j, t_j]$ from value ranking;
4. Define logical OR: $I = \bigcup_{j=1}^{N_{FR}} [s_j, t_j]$;
5. Find hyper-rectangle
$P(\Delta) = \bigcup_{j=1}^{N_{FR}} [s_j \mp \delta_{s_j} \cdot s_j, t_j \pm \delta_{t_j} \cdot t_j]$;
6. Find optimal perturbations Δ^*

Both methods consist in determining the optimal shape of an hyper-rectangle $\mathcal{P}(\Delta)$, by finding the optimal perturbations Δ^* of value ranking thresholds.

This can be differently formalized in the following way.

Reliability from outside starts from the unsafe class, $y = 1$ in our case. The solution is found by solving:

$$\Delta^* = arg \min_{\Delta: N_1 = D_1} \mathcal{V}(\mathcal{P}(\Delta)) \tag{1}$$

being N_1 the number of elements in $\mathcal{X}$ classified as $y = 1$ and included into hyper-rectangle $\mathcal{P}$ with volume $\mathcal{V}$. Since the optimal $\mathcal{P}$ contains all unsafe points, the complementary is considered as safety region.

Reliability from inside starts from the safe class ($y = 0$). The optimal solution is now as follows:

$$\Delta^* = arg \max_{\Delta: N_1 = 0}, \mathcal{V}(\mathcal{P}(\Delta)) \tag{2}$$

and $\mathcal{P}(\Delta^*)$ is the safety region containing no unsafe points. Hyper-rectangles might be too simple to follow potentially complex boundaries between output classes, so an alternative is *LLM0%* method. The m^0 highest covering rules obtained for the safe class by training the LLM with 0% maximum error are joined in logical OR, obtaining predictor $\hat{r}$. Feature ranking can be exploited to select N_{FR} features and apply perturbations $\boldsymbol{\delta}$ on their thresholds contained in $\hat{r}$, thus having $\hat{r}(\boldsymbol{\delta})$. The optimal perturbations are chosen according to the following problem:

$$\boldsymbol{\delta}^* = arg \max_{\boldsymbol{\delta}: E(\hat{r}(\boldsymbol{\delta}))=0} C(\hat{r}(\boldsymbol{\delta})) \tag{3}$$

2.3 Safe SVDD

The search for totally safe regions is an open and remarkable problem in ML. In our studies, we proposed a method for the research of zero FPR (False Positive Rate) regions [Carlevaro e Mongelli, 2021b; Carlevaro e Mongelli, 2021a]: it performs successive iterations of the SVDD [Tax e Duin, 1999] on a safe initial region, found with a preliminary SVDD, until there are no more negative points inside it. We achieve convergence when we reach a fixed number of iterations or when the condition on FPR is satisfied.

Algorithm 1 `ZeroFPRSVDD`
Data set $\mathcal{X}\times\mathcal{Y}$ is divided in training set $\mathcal{X}_{tr}\times\mathcal{Y}_{tr}$ and test set $\mathcal{X}_{ts}\times\mathcal{Y}_{ts}$. A threshold of ε is set.

1. SVDD-**cross-validation** on $\mathcal{X}_{tr}\times\mathcal{Y}_{tr}$
2. S_1=SVDD($\mathcal{X}_{tr}, \mathcal{Y}_{tr}, C_{-1}, C_{+1}, \sigma$)
3. **Test** SVDD on $\mathcal{X}_{ts}\times\mathcal{Y}_{ts}$
4. maxiter=1000;
5. i=2;
6. **while**(i<maxiter)
6.1. $\mathcal{X}_{tr_i} = S_i(\mathcal{X}_{ts})$;
6.2. SVDD-**cross-validation** on $\mathcal{X}_{tr_i}\times\mathcal{Y}_{tr_i}$
6.3. S_i=SVDD($\mathcal{X}_{tr_i}, \mathcal{Y}_{tr_i}, C_{-1}, C_{+1}, \sigma$)
6.4. **Test** SVDD on $\mathcal{X}_{ts}\times\mathcal{Y}_{ts}$
6.5. **if**(FPR< ε)
6.5.1. **return** $S^* = S_i$;
6.5. **end**
7. $i = i + 1$;
end

2.4 Learning Assurance in XAI

Generalization

Generalization is the ability of a learning algorithm to perform with acceptable levels on new data, unseen during the training phase, but coming from the same data distribution as the one used in model design.

To quantify it, different techniques have been developed so far in statistical learning theory, by finding a *generalization gap*, or *bound,* representing the difference between true (ideal) and empirical risk, in probabilistic terms.

Most of the existing bounds require huge amounts of data or a measure of the model complexity (e.g. VC-dimension) [Tempo *et al.*, 2013].

However, Clopper-Pearson (CP) bound can be empirically evaluated on a test set, hence it can be used in practice to assess LLM generalization capabilities [Wolf, 2018]. Consider a function h, known as the *hypothesis*, mapping input $x \in \mathcal{X}$ into an output class $y \in \mathcal{Y}$. Also, let $R(h)$ be the true risk of the learning algorithm. Then, with probability at least $1 - \epsilon$ over an i.i.d. draw of a test set $T \in (\mathcal{X} \times \mathcal{Y})^m$, the CP bound is found as:

$$R(h) \leq \hat{R}_T(h) + \sqrt{\frac{\ln \frac{1}{\epsilon}}{2m}}, \tag{4}$$

$\hat{R}_T(h)$ is the empirical error over T:

$$\hat{R}_T(h) = \frac{1}{|T|} \sum_{(x,y)\in T} (\mathcal{L}(y, h(x))), \tag{5}$$

and $\mathcal{L}$ being the 0-1 loss function.

Moreover, a solution independent from the model complexity is also proposed by our group in [Mirasierra *et al.*, 2022]. It studies less conservative bounds than Clopper-Pearson as the bounds on the sample-complexity of the validation dataset are specific to the given algorithm/dataset, thus allowing to drastically reduce conservatism (and hence optimize data usage).

Robustness

While generalization assumes that a test set is drawn from the same distribution of the training data, a trustworthy model should also behave correctly when data belong to different distributions. And this is when *robustness* comes into play. In general, a model is assumed to work in given ranges, known as the nominal work domain, defined during the training phase. Robustness is achieved when the model continues

Authors	Reference	Short Abstract	Methodology	Strengths
Carlevaro et al.	*A New SVDD Approach to Reliable and eXplainable AI* [Carlevaro e Mongelli, 2021a]	The works address how SVDD can be redesigned to detect safety regions in a cyber physical system with zero statistical error. Rule-based knowledge extraction is also presented, to let the SVDD be understandable	Sections (2.1), (2.3)	Detection of regions with (almost) zero statistical error in classification, extraction of rules from SVDD.
Vaccari et al.	*eXplainable and Reliable against adversarial machine learning in data analytics* [submitted]	In this work a new approach able to detect and mitigate malicious adversarial machine learning attacks against a machine learning system is proposed. Three methodologies to enhance reliability of the detection are discussed and presented, being based on an explainable-by-design algorithm.	Sections (2.1), (2.3), (2.2)	Collection of detection algorithms against adversary ML, rule extraction algorithms.
Narteni et al.	*From Explainable to Reliable Artificial Intelligence [Narteni* et al.*, 2021]*	Three methods to design safety regions with zero error, based on native XAI and its properties, are introduced and tested with promising results in safety-critical applications.	Sections (2.1), (2.2)	Being entirely based on XAI, the interpretation of the results is straightforward. Also, the methods could also be combined to further enhance the performance and adapt to the data at hand.
Lenatti et al.	*Counterfactual building and evaluation via eXplainable Support Vector Data Description* [submitted]	The paper addresses counterfactuals through SVDD, empowered by explainability and metric for assessing the counterfactual quality.	Section (2.5)	Innovative method for extracting counterfactuals, deep insight into counterfactual literature.
Dabbene et al.	*Prediction error quantification through probabilistic scaling* [Mirasierra *et al.*, 2022]	This study addresses the quantification of the probabilistic error of prediction models by means of a probabilistic scaling method.	Section (2.4)	The required number of observations is independent of the complexity of the prediction model.
Muselli et al.	*Switching Neural Networks: A New Connectionist Model for Classification [Muselli, 2006]*	Foundation of the Logic Learning Machine.	Section (2.1)	A new approach to eXplainable AI.
Mongelli et al.	*Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence [Mongelli, 2021]*	Sensitivity analysis of eXplainable AI for safety in cyber physical systems.	Sections (2.1), (2.2)	A new approach in the intersection of machine learning and safety.
Narteni et al.	*Trustworthy Classification-based Equivalent Bandwidth Control*, [submitted]	A unified view of machine learning and control, where XAI is used to derive insights into the structure of the problem by means of intelligible rules. The study also identifies methods to test generalization and robustness of the proposed model.	Section (2.4)	The adoption of XAI provides innovative solutions to generalization and robustness.

Table 1: Overview of related literature.

to operate correctly even outside of this domain. Hence, robustness tests may consist in applying the model over larger ranges of the input features than those used in training. It is crucial to understand *when* the operational conditions are different from the nominal ones. This would lead to understand if the samples are *in-distribution* or *out-of-distribution*, namely, if the system is working (at the operational level) under the same probability distribution used in training. Since such probability distribution may be hardly derived from data, the in- versus out-of- distribution question constitutes one of the most challenging issues in machine learning today.

XAI facilitates approaching the problem. Each sample may verify or not any rule provided by the LLM: a histogram representing the number of times each rule is verified can give useful information. The changes of the histograms with respect to the baseline may give evidence of out-of-distribution samples. The metrics used to quantify those changes can be simple statistics such as mean, variance, skewness and kurtosis, as well as the mutual information between several couples of histograms. Once these metrics differ from the baseline, an out-of-distribution alarm may be consequently triggered.

2.5 Counterfactual explanations

Given a binary classification problem described by the class S_1 and S_2, the counterfactual explanation (CE) of a certain observation $\boldsymbol{x} \in S_1$ is the minimum changes in the input features that lead to the generation of another observation $\boldsymbol{x}^* \in S_2$. Counterfactual explanation falls in the class of local explainability techniques as they give information about the logic behind the prediction of individual observations.

We recently introduced a novel methodology for constrained counterfactual generation and validation, based on minimal perturbations of controllable features $\mathbf{u}$, maintaining fixed the non controllable ones $\mathbf{z}$ (in our formulation, $\boldsymbol{x} = (\mathbf{u}, \mathbf{z})$). The counterfactual generation method, whose numerical solution is summarized by **Algorithm 3**, uses regions defined by Two Class-Support Vector Data Description (TC-SVDD). Under simplified conditions (Euclidean distance and linear kernel), an analytical solution of the counterfactualan analytical solution of the optimal counterfactual (in terms of minimum distance) is shown as well. The validation method combines computational simulations and XAI, exploited in terms of readily interpretable rule-based classification with LLM models, allowing for a robust validation of correctness and consistency of the generated explanations.

Algorithm 3 `CounterfactualSVDD`
Dataset $\mathcal{X} \times \mathcal{Y} \subset R^N \times \{-1, +1\}$ is divided in training set $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validation set $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$.
A TC-SVDD is performed on $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validated on $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$ in order to derive S_1 and S_2.
$N_{\text{grid}}, N_{\text{cntrfctls}} > 0$ are fixed.

---

1. $\mathcal{C} = [\,]$
2. Build a $(N_{\text{grid}} \times N_{\text{grid}})^N$ grid G
3. $G_1 \cup G_2 \doteq G \cap \big(S_1 \triangle S_2\big)$
4. **for** $i = 1 : N_{\text{cntrfctls}}$
4.1 $x_i{=}(\mathbf{u}_i, \mathbf{z}_i){\in}S_1$
4.2 $d_i = d\big(x_i, G_{2_{|\mathbf{z}=\mathbf{z_i}}}\big)$
4.3 $x'_i = \min(d_i)$
4.4 **if**$\big(x_i \in S_1$ & $x'_i \in S_2\big)$
4.4.1 $\mathcal{C} = \mathcal{C} \cup \{\boldsymbol{x}'_i\}$
4.5 **end**
5. **end**
6. **return** $\mathcal{C}$

---

3 Scientific Contributions

The research carried out by our team covers several fields of reliability in ML and, as a result, a large amount of literature has been produced in all scientific areas mentioned so far. The table 1 summarizes some of our work, that which we think best highlights our contribution to the scientific community. In addition, our team has been involved in various research projects, covering both theoretical and applied issues. Below, we would like to summarise the projects that we feel have contributed most to the progress of our group and the scientific community:

- *Genova 5G*: tender on safety of road infrastructures in the territorial area of Genova through experiments based on Vodafone 5G technology, Ministry of Economic Development (MiSE), November 1, 2020 - July 31, 2022;
- *SafeCOP ('Safe Cooperating Cyber-Physical Systems using Wireless Communication')*[2]: study and implementation of certifiable interconnections of cyber physical systems; it includes machine learning for performance prediction in new generation intelligent transport systems ('vehicle to infrastructure' and 'vehicle to vehicle'), call EU ECSEL 2016, April 15, 2017 - July 31, 2019;
- *GESTEC('Service-oriented technologies for the development and integration of ICT platforms')*: study of machine learning, cybersecurity, and Software Defined Networking techniques in the cloud environment, Ministry of Education, University and Research (MIUR), July 15, 2018;
- Participation in *SAE G-34/EUROCAE WG-114 AI in Aviation Committee*[3]

[2] www.safecop.eu

References

[Carlevaro e Mongelli, 2021a] A. Carlevaro e M. Mongelli. A new svdd approach to reliable and explainable ai. *IEEE Intelligent Systems*, (01):1–1, oct 2021.

[Carlevaro e Mongelli, 2021b] A. Carlevaro e M. Mongelli. Reliable ai trough svdd and rule extraction. *International IFIP Cross Domain (CD) Conference for Machine Learning & Knowledge Extraction (MAKE), CD-MAKE 2021.*, 2021.

[Guidotti *et al.*, 2019] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, e Dino Pedreschi. A survey of methods for explaining black box models, Sep 2019.

[Mirasierra *et al.*, 2022] Victor Mirasierra, Martina Mammarella, Fabrizio Dabbene, e Teodoro Alamo. Prediction error quantification through probabilistic scaling. *IEEE Control Systems Letters*, 6:1118–1123, 2022.

[Mongelli, 2021] M. Mongelli. Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence. *Computer Communications*, 2021.

[Muselli, 2006] Marco Muselli. Switching neural networks: A new connectionist model for classification. In Bruno Apolloni, Maria Marinaro, Giuseppe Nicosia, e Roberto Tagliaferri, editors, *Neural Nets*, pages 23–30, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[Narteni *et al.*, 2021] S. Narteni, M. Ferretti, V. Orani, I. Vaccari, E. Cambiaso, e M. Mongelli. From explainable to reliable artificial intelligence. *International IFIP Cross Domain (CD) Conference for Machine Learning & Knowledge Extraction (MAKE), CD-MAKE 2021.*, 2021.

[Tax e Duin, 1999] D. Tax e R. Duin. Support vector domain description. *Pattern Recognition Letters 20*, pages 1191–1199, 1999.

[Tempo *et al.*, 2013] Roberto Tempo, Giuseppe Calafiore, e Fabrizio Dabbene. Statistical learning theory. In *Randomized Algorithms for Analysis and Control of Uncertain Systems*, pages 123–134. Springer, 2013.

[Wolf, 2018] Michael M Wolf. Mathematical foundations of supervised learning, 2018.

[3] https://www.eurocae.net/news/posts/2019/june/new-working-group-wg-114-artificial-intelligence/