This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

Sensitivity of Logic Learning Machine for reliability in safety-critical systems

Sara Narteni
Politecnico di Torino, CNR-IEIIT

Vanessa Orani
CNR-IEIIT

Ivan Vaccari
CNR-IEIIT

Enrico Cambiaso
CNR-IEIIT

Maurizio Mongelli
CNR-IEIIT

***Abstract*—Nowadays, Artificial Intelligence (AI) is bursting in many fields, including critical ones, giving rise to reliable AI, that means ensuring safety of autonomous decisions. As the false negatives may have a safety impact (e.g., in a mobility scenario, prediction of no collision, but collision in reality), the aim is to push false negatives as close to zero as possible, thus designing "safety regions" in the feature space with statistical zero error. We show here how sensitivity analysis of an eXplainable AI model drives such statistical assurance. We test and compare the proposed algorithms on two different datasets (physical fatigue and vehicle platooning) and achieve quite different conclusions in terms of results that strongly depend on the level of noise in the dataset rather than on the algorithms at hand.**

THE INTRODUCTION. The rapid growing of Artificial Intelligence (AI) technologies is leading to a closer interaction between humans and machines, which poses legal and ethical issues; technology experts and policy makers should cooperate in order to make AI trustworthy and responsible [1]. To this effort, regulation is being developed, stating the requirements that AI systems should follow to achieve such goals. For example, the European GDPR (https://gdpr.eu/tag/gdpr/), introduced in 2018, states that a *"right to explanation"* is needed when dealing with automated systems. This has paved the way to the development of a subfield of AI, referred to as eXplainable AI (XAI), aiming to provide humans with understanding and trust in models outcomes. XAI may come in the form of global intelligible rules, being simpler and generally less accurate than more sophisticated models (like those of deep learning), but with the advantage of being interpretable.

Another viewpoint to trustworthy AI is identifying and handling assurance under uncertainties in AI systems. This means improving reliability of prediction confidence. The topic remains a significant challenge in machine learning (ML),

as learning algorithms proliferate into difficult real-world pattern recognition applications. The intrinsic statistical error introduced by any ML algorithm may lead to criticism by safety engineers. This is corroborated even more by the intrinsic instability of deep learning (DL) in the presence of malicious noise [2]. The topic has received a great interest from industry [3], in particular in the automotive [4] and avionics [5] sectors. In this context, the conformal predictions framework [6] studies methodologies to associate reliable measures of confidence with pattern recognition settings, including classification.The proposed approaches follows this direction, by identifying methods to design data-driven safety envelopes with statistical zero error. We show how this assurance may limit the size of the safety envelope (e.g., collision avoidance by drastically reducing vehicles speed) and focus on how to find a good balance between the assurance and the extension of working conditions.

To this aim, we study how XAI can achieve reliability based on rule tuning around safety regions. The rationale behind the choice of XAI relies on the typical complex shape of the solutions drawn by black-box techniques (such as neural networks), which makes a similar sensitivity analysis a hard task. We study the certification under different angles of the estimated safety envelope. Moreover, by testing and comparing our three proposed methodologies on two distinct datasets, we highlight the alternatives which can be tuned according to the characteristics of the dataset under analysis.

Related Work

Nowadays, a big effort is put in research to integrate ML algorithms with safety, since erroneous predictions may lead to severe consequences in many safety-critical fields [7]. In the context of autonomous driving, typical safety engineering approaches are considered, with extension to ML paradigm [8]. These certification approaches include formal verification [9], transparent implementation [10], uncertainty estimation [11], error detection [12], domain generalization [13] and adversarial approaches based on data perturbation and corruption [14].

Other methods integrate safety assurance into reinforcement learning, by making predictions to guide the agent towards safe decisions [15].

Integrating ML and DL systems with safety and reliability is fundamental in healthcare too: in [16] a method to assess safety for pattern recognition using a medical device is developed.To the best of our knowledge, very few studies address ML trustworthiness based on explainable AI. Table 1 summarizes how the present work contributes to advance the presented state-of-the-art with respect to different key points: nature of the algorithms (XAI or DL), kind of the data (images or other) and safety approach (error control or perturbation control).

Logic Learning Machine

Logic Learning Machine (LLM) is a global rule-based method, developed as an improvement of Switching Neural Networks [17] by Rulex (https://www.rulex.ai/rulex-explainable-ai-xai/). The LLM builds a classifier $g(x)$ described by rules of the following format: **if** $<$*premise*$>$ **then** $<$*consequence*$>$. The $<$*premise*$>$ is the logical AND ($\wedge$) of conditions on the input features, whereas $<$*consequence*$>$ is the output class. The model is built in three phases:

1) *Discretization and Latticization*: each variable is transformed into a string of binary data in a Boolean lattice, using the inverse only-one code binarization. All the strings are then concatenated in one large string per each sample.
2) *Shadow Clustering*: a set of binary values, called implicants, are generated, allowing the identification of groups of points associated with a specific class.
3) *Rule Generation*: the implicants are transformed into simple conditions and combined into a collection of intelligible rules.

An implicant is a binary string in a Boolean lattice that uniquely determines a group of points associated with a given class. Starting from an implicant, it's possible to derive a rule having in its premise a logical product of conditions based on cutoffs obtained during the discretization step. In LLM all the implicants are generated via Shadow Clustering by looking at the whole training set: thus, resulting rules can overlap and represent different relevant aspects of the underlying problem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

	eXplainable AI	Deep Learning	Image data	Multivariate data	Error control	Perturbation control
[2]		×	×			×
[4]			×	×		
[5]		×	×		×	
[6]				×	×	
[7]	×				×	
[8]					×	
[9]		×	×	×		×
[10]	×	×	×			
[11]		×	×			×
[12]		×	×	×	×	
[13]		×	×			
[14]		×	×			×
[15]				×	×	
[16]		×		×	×	
This paper	×			×	×	

Table 1: The table summarizes the positioning of the state of the art with respect to the main pillars of trustworthy AI. Error control here denotes additional mechanisms to give statistical guarantees to the model. Perturbation control means robustness to discrepancies between training and operational data

.

Feature and Value Ranking

A XAI model allows to perform an inspection of its outcomes through feature and value ranking.

Consider a set of m rules $\mathbf{r}_k, k = 1, \ldots, m$, each including d_k conditions $c_{l_k}, l_k = 1_k, \ldots, d_k$. Let $X_1, \ldots, X_n$ be the input variables, s.t. $X_j = x_j \in \mathcal{X} \subseteq \mathbb{R} \quad \forall j = 1, \ldots, n$. Let also $\hat{y}$ be the class assigned by the rule and y_j the real output of the $j-th$ instance.

A condition c_{l_k} involving the variable X_j, can assume one of the following forms:

$$X_j > s, \quad X_j \leq t, \quad s < X_j \leq t, \qquad (1)$$

being $s, t \in \mathcal{X}$.

For each rule, it is possible to define a confusion matrix. It is made up of four indices: $TP(\mathbf{r}_k)$ and $FP(\mathbf{r}_k)$, defined as the number of instances (x_j, y_j) that satisfy all the conditions in rule $\mathbf{r}_k$ with $\hat{y} = y_j$ and $\hat{y} \neq y_j$ respectively; $TN(\mathbf{r}_k)$ and $FN(\mathbf{r}_k)$, defined as the number of examples (x_j, y_j) which do not satisfy at least one condition in rule $\mathbf{r}_k$, with $\hat{y} \neq y_j$ and $\hat{y} = y_j$, respectively.

Consequently, we can derive the following metrics:

$$C(\mathbf{r}_k) = \frac{TP(\mathbf{r}_k)}{TP(\mathbf{r}_k) + FN(\mathbf{r}_k)} \qquad (2)$$

$$E(\mathbf{r}_k) = \frac{FP(\mathbf{r}_k)}{TN(\mathbf{r}_k) + FP(\mathbf{r}_k)} \qquad (3)$$

The covering $C(\mathbf{r}_k)$ is adopted as a measure of relevance for a rule $\mathbf{r}_k$; in other words, the greater is the covering, the higher is the generality of the corresponding rule. The error $E(\mathbf{r}_k)$ is a measure of how many data are wrongly covered by the rule. Both covering and error are used to define feature ranking and value ranking.

Feature ranking (FR) provides a ranking of the features used into the rules conditions according to a measure of relevance. In order to obtain the relevance $R(c_{l_k})$ for a condition, we consider rule $\mathbf{r}_k$, in which condition c_{l_k} occurs, and the same rule without condition c_{l_k}, denoted as $\mathbf{r}'_k$. Since the premise part of $\mathbf{r}'_k$ is less stringent, we obtain that $E(\mathbf{r}'_k) \geq E(\mathbf{r}_k)$, thus the quantity $R(c_{l_k}) = (E(\mathbf{r}'_k) – E(\mathbf{r}_k))C(\mathbf{r}_k)$ can be used as a measure of relevance for the condition of interest c_{l_k}. Each condition c_{l_k} refers to a specific variable X_j and is verified by some values $\nu_j \in \mathcal{X}$. In this way, a measure of relevance $R_{\hat{y}}(\nu_j)$ for every value assumed by X_j is derived by the following equation 4:

$$R_{\hat{y}}(\nu_j) = 1 - \prod_k \left(1 - R\left(c_{l_k}\right)\right), \qquad (4)$$

where the product is computed on the rules $\mathbf{r}_k$ that include a condition c_{l_k} verified when $X_j = \nu_j$. Since $R_{\hat{y}}(\nu_j)$ takes values in $[0, 1]$, it can be interpreted as the probability that value ν_j occurs to predict $\hat{y}$. The same argument can be extended to intervals $I \subseteq \mathcal{X}$, thus defining the *Value Ranking (VR)*. Relevance scores are then ordered, giving evidence of the most sensitive

interval of the feature with respect to each class.

Reliability Assessment Methods

Considering a binary classification problem, we refer to the positive class ($y = 1$) as the unsafe one. In contrast, class $y = 0$ is referred to as the safe class. Based on this, we call "safety regions" the regions in the feature space where false negatives tend to zero; that means the prediction of safety has no statistical error. Through LLM feature and value ranking, we apply two sensitivity optimization problems to shape such regions. Moreover, the LLM itself may be posed with statistical zero error and be used in comparison with the former approaches. The three methodologies are detailed formulated in the following.

Reliability from Outside

Let X be a $D \times N$ matrix of the input vectors $x_i \in \mathbb{R}^N$, with N being the total number of features and $i \in [1, D]$. Let $g(x_i) = y$ be the function describing the LLM classification (hence, $g(x_i) = 1$ for the positive class). Let D_1 be the number of instances belonging to class $y = 1$ and D_0 the number of instances in class $y = 0$, so that $D_1 + D_0 = D$.

Let N^{FR} be the number of the most significant features obtained through the feature ranking for class $y = 1$. For each feature $j \in [1, N^{FR}]$, we can use the LLM value ranking to define the most significant interval for $y = 1$ as $[s_j, t_j]$. Our method consists in expanding it as follows: $[s_j - \delta_{s_j} \cdot s_j, t_j + \delta_{t_j} \cdot t_j]$.

Being $\Delta = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{N^{FR}})$ a matrix, with $\boldsymbol{\delta}_j = (\delta_{s_j}, \delta_{t_j})$, the optimal Δ is computed through the following optimization problem. Let $\mathcal{P}(\Delta)$ be the hyper-rectangle under the expanded intervals and let $\mathcal{V}(\mathcal{P}(\Delta))$ be the inherent volume.

Then, the optimization problem identifies the best fit from the outside of class $y = 1$, namely, it finds the most suitable shape, in terms of rule-based intervals, of safe points around the unsafe ones. It is as follows:

$$\Delta^* = arg \min_{\Delta : N_1 = D_1} \mathcal{V}(\mathcal{P}(\Delta)) \tag{5}$$

being N_1 the number of elements in X classified as $y = 1$ and included into $\mathcal{V}(\mathcal{P}(\Delta))$.

For instance, if we fix N^{FR}=2, the hyper-rectangle $\mathcal{P}$ becomes a rectangle $\mathcal{S}$. The optimization process lets us find out the matrix $\Delta^* = (\boldsymbol{\delta}_1^*, \boldsymbol{\delta}_2^*)$. The related optimal intervals are $I_1 = (s_1 - \delta_{s_1}^* \cdot s_1, t_1 + \delta_{t_1}^* \cdot t_1), I_2 = (s_2 - \delta_{s_2}^* \cdot s_2, t_2 + \delta_{t_2}^* \cdot t_2)$, corresponding to the features $j = 1$ and $j = 2$ respectively: their logical union ($\vee$) defines a surface $\mathcal{S}$.

Then, the "safety region" is defined as the complementary bi-dimensional surface of $\mathcal{S}$, which can be written as follows:

$$\begin{aligned} \mathcal{S}_1 = ((-\infty, s_1 - \delta_{s_1}^* \cdot s_1) \vee (t_1 + \delta_{t_1}^* \cdot t_1, \infty)) \wedge \\ ((-\infty, s_2 - \delta_{s_2}^* \cdot s_2) \vee (t_2 + \delta_{t_2}^* \cdot t_2, \infty)) \end{aligned} \tag{6}$$

Reliability from Inside

This method performs the same search for "safety regions", but it starts with N^{FR} most important features for safe class instead and reduces their most relevant intervals (again, provided by LLM value ranking) until the obtained region only contains true negative instances.

In this case, the reduced intervals are: $[s_j + \delta_{s_j} \cdot s_j, t_j - \delta_{t_j} \cdot t_j]$. Being Δ defined in the same way as for Equation 5 and $\mathcal{P}_0$ the hyper-rectangle under the reduced intervals, the optimal Δ is found by enlarging as much as possible the hyper-rectangle from inside the safe class, until an unsafe point is reached. It is as follows:

$$\Delta^* = arg \max_{\Delta : N_1 = 0} \mathcal{V}(\mathcal{P}_0(\Delta)) \tag{7}$$

For $N^{FR} = 2$, the "safety region" is the following surface $\mathcal{S}_0$:

$$\begin{aligned} \mathcal{S}_0 = (s_1 + \delta_{s_1}^* \cdot s_1, t_1 - \delta_{t_1}^* \cdot t_1) \vee \\ (s_2 + \delta_{s_2}^* \cdot s_2, t_2 - \delta_{t_2}^* \cdot t_2) \end{aligned} \tag{8}$$

LLM with Zero Error

As the sharp angularity of hyper-rectangles may be not fine enough to follow the potential complex shapes of the boundaries between the classes, a more refined approach would ask for more complex separators, still preserving the zero error constraint and by starting from the available rule baseline. Zero error classification (for the safe class) is readily available by the shadow clustering adopted by LLM. The clustering process

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

is applied with the further constraint of building clusters without superposition of points of more than one class [18] (LLM 0%, in the following). All the resulting rules with zero error are then joined in logical OR ($\vee$), thus describing a shape more complex than a hyper-rectangle. The new model deserves a further sensitivity tuning (on a test set) as follows.

The LLM 0% defines a set of m rules $\mathbf{r}_k,\ k = 1,\ldots,m$ so that $E(\mathbf{r}_k) = 0\ \forall k \in [1,m]$. Suppose that this provides a set of m^0 rules $\mathbf{r}_k^0,\ k = 1,\ldots,m^0$ for the safe class ($y = 0$). Also, let $c_{l_k}^0,\ l_k^0 = (1,\ldots,d_k^0)$ be the set of d_k^0 conditions inside of each rule $\mathbf{r}_k^0$. Then, we consider the logical OR ($\vee$) between the obtained highest-covering rules, building a new predictor $\hat{r}$. Our goal is to assess its ability of classifying new test set data with statistical zero error (FNR=0). This implies to further tune $\hat{r}$, by reducing a subset of its conditions $c_{l_k}^0$, chosen as those containing the first N^{FR} features obtained from LLM 0% feature ranking for class $y = 0$. In mathematical terms, for each feature $j \in [1, N^{FR}]$, we add the thresholds of the chosen conditions by applying $\boldsymbol{\delta} = (\delta_s, \delta_t)$, being δ_s and δ_t the perturbations applied to s and t thresholds, respectively, as defined in Equation 1. Let $\hat{r}(\boldsymbol{\delta})$ be the resulting perturbed predictor, our goal is then to find the optimal $\boldsymbol{\delta}$ as follows:

$$\boldsymbol{\delta}^* = arg \max_{\boldsymbol{\delta}: E(\hat{r}(\boldsymbol{\delta}))=0} C(\hat{r}(\boldsymbol{\delta})) \qquad (9)$$

Applications and Results

The methods described in the previous section have been applied on two different classification problems: physical fatigue detection in working task simulation and collision detection in vehicle platooning. Together with the false negative rate (FNR), the true negative rate (TNR) is of interest, being the measure of the quantity of safe points in a region. The aim is to obtain the largest TNR with zero FNR.

Physical Fatigue

The data used in this application belong to an open-source dataset (https://github.com/zahrame/FatigueManagement.github.io/tree/master/Data). Data were collected through wearable sensors, i.e. Inertial Movement Units, from 15 participants who were asked to perform a simulation of an industrial task for 180 minutes and provide a fatigue level every 10 minutes using RPE [19]. A RPE$\geq$13 corresponds to a fatigued state (class $y = 1$), otherwise to non-fatigued (class $y = 0$). From sensors raw data, a list of features is derived (see Table 2 in [20]). We removed heart-rate related features as well as gender, since it is not numerical, and standardized data by applying z-score transformation.

We then trained LLM with standard 5% maximum error allowed for rules on a 67% training set. We evaluated it on a 33% test set obtaining accuracy of 82%, sensitivity of 71%, specificity of 95% and F1-score of 81 %.

Reliability from Outside

To test this method, we considered the first two most important intervals for fatigued class that we got from LLM value ranking: *back rotation position in sagittal plane* $>$ 0.03 and *wrist jerk coefficient of variation* $>$ 0.03. We applied the optimization algorithm (Equation 5) on such intervals and obtained $\delta_{s_1}^* = -13, \delta_{s_2}^* = 28$, resulting in FNR=0 and TNR=0.20. Therefore, the "safety region", which we call "non-fatigue region" in this context, can be expressed as follows (for brevity, let f_1 and f_2 be the two above mentioned features):

$$\mathcal{S}_1 = ((f_1 \in (-\infty, 0.42)) \wedge (f_2 \in (-\infty, -0.81))$$

The region was then validated in order to take into account that the involved feature values should vary in a limited range, so to reflect real human movement capabilities and correspond to proper execution of the task. Since the dataset documentation does not drive in this direction and the inherent literature lacks of standard ranges, we chose to consider maximum and minimum values for the features based on two age groups (age$\leq$40 and age$>$40). This helps to highlight the further stratification readily available from the sensitivity analysis.

Doing so, we were able to redefine two "non-fatigue regions" by limiting the previous one according to the ranges we found; such new regions are expressed as:

$$\mathcal{S}_1 = ((f_1 \in (-2.52, 0.42)) \wedge (f_2 \in (-1.78, -0.81)) \text{ for } age \leq 40 \text{ y.o.}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

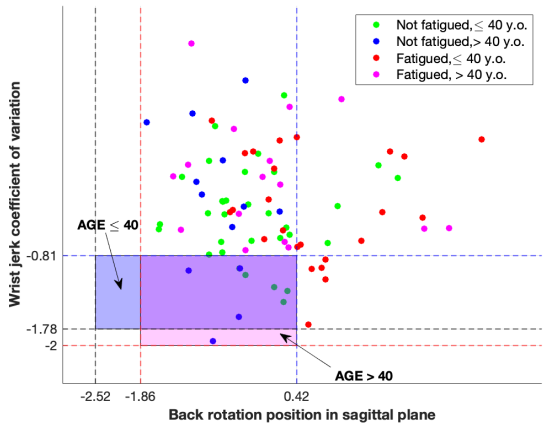

Figure 1: Scatter plot of the first two features (back rotation position in sagittal plane and wrist jerk coefficient of variation) with representations of the "non-fatigue region" (FNR=0) individuated for age $\leq$ 40 group (pink) and age $>$ 40 (violet).

$$\mathcal{S}_1 = ((f_1 \in (-1.86, 0.42)) \wedge (f_2 \in (-2.0, -0.81)) \text{ for } age > 40 \text{ y.o.}$$

In Fig. 1 a visual representation of the obtained regions is provided.

Reliability from Inside

Reliability from inside considered the problem of identifying non-fatigue regions starting from the non-fatigued class The value ranking shown *back rotation position in sagittal plane* $\leq$ 0.03 and *chest acceleration mean* $>$ -0.47 as the two most relevant intervals for predicting non-fatigue. On such conditions, we applied the optimization problem (Equation 7), which led us to individuate $\delta^*_{t_1} = 57$, $\delta^*_{s_2} = 8.78$. For these values, we got FNR=0 and TNR=0.06. The "non-fatigued region" $\mathcal{S}_0$ is then found (with f_1 and f_2 being *back rotation position in sagittal plane* and *chest acceleration mean* respectively):

$$\mathcal{S}_0 = (f_1 \in (-\infty, -1.68) \vee f2 \in (3.65, \infty))$$

Just as for the outside approach, we limited such region in function of the two age groups ($\leq$ and $>$ 40 years old). This redefines $\mathcal{S}_0$ as follows (see Fig. 2 for the graphical representation):

$$\mathcal{S}_0 = (f_1 \in (-2.52, -1.68) \vee f2 \in (3.65, 3.99)) \; for \; age \leq 40 \; y.o.$$

$$\mathcal{S}_0 = (f_1 \in (-1.86, -1.68) \vee f2 \in (3.65, 3.99)) \; for \; age > 40 \; y.o.$$

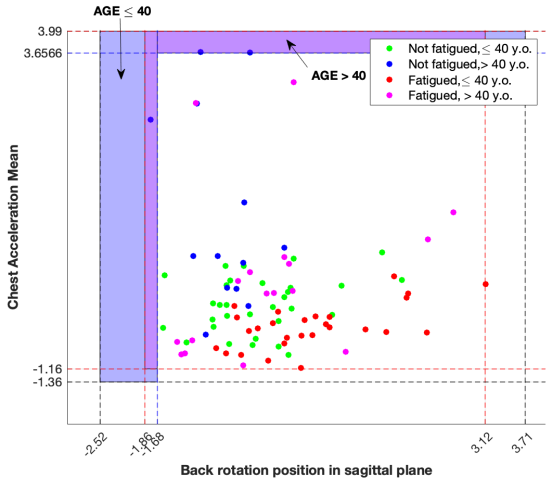

Figure 2: Scatter plot of the first two features (back rotation position in sagittal plane, Chest Acceleration Mean) from value ranking of non-fatigued class, with representations of the "non-fatigue regions" (FNR=0) based on the age group (violet for age $\leq$ 40, pink otherwise)

Zero Error LLM

Both previous approaches have the limitation of individuating optimal solutions to the identification of "non-fatigue regions" characterized by relatively low values of TNR, i.e. number of points included in such surfaces.

To assess if such values could be increased, we trained the LLM 0%, obtaining 6 rules for the safe class, and built a new predictor by joining the first four highest-coverage rules in logical OR (see below).

IF (0.51 $<$ HipACCMean $\leq$ 1.98 **and**
ChestACCcoefficientofvariation $\leq$ 1.11 **and** -1.73 $<$ averagestepdistance $\leq$ 0.81 **and** backrotationpositioninsagplane $\leq$ 0.52) $\vee$
(WristjerkMean $>$ 0.55 **and** -1.35 $<$ Back rotation position in sag plane $\leq$ 0.04) $\vee$
(-1.73 $<$ averagestepdistance $\leq$ -0.22 **and** backrotationpositioninsagplane $\leq$ -0.25 **and** -0.44 $<$ numberofsteps $\leq$ 3.75 **and** -1.73 $<$ Wristjerkcoefficientofvariation $\leq$ 0.55) $\vee$
(ChestxpostureMean $>$ -0.033 **and** HipzpostureMean $>$ 0.43 **and** WristACCMean $>$ -0.83 **and** -0.88 $<$ backrotationpositioninsagplane $\leq$ 0.29)
THEN ***non-fatigued***

Prior to any perturbation, we got FNR=0.06 and TNR=0.75. To further decrease the FNR, we conducted the optimization process described in Equation 9 by tuning the thresholds for the first $N^{FR} = 2$ features from non-fatigued feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

ranking, namely *HipACCMean* and *WristjerkMean*. We obtained $\delta^*_{s_1}$ =1.848 and $\delta^*_{t_2}$ =0.027 for such features respectively, with FNR=0 and TNR=0.42.

Vehicle Platooning

Vehicle platooning is one of the most important challenges in autonomous driving, dealing with a trade-off between performance and safety. In our analysis we considered a scenario of cooperative adaptive cruise control as described in [18], where the platoon is in a steady state of speed and reciprocal inter-vehicular distance when a braking force is applied by the leader of the platoon. In the present application, we used simulation data generated by Plexe simulator (https://github.com/mopamopa/Platooning). For each of the generated samples, 5 features were computed and filtered within the following ranges: the number of vehicles, $N \in [3, 8]$ the braking force $F_0 \in [-8, -1] \times 10^3$ N the Packet Error Rate $PER \in [0.2, 0.5]$ the initial distance between vehicles $d(0) \in [4, 9]$ m (supposed equal for all of them); the initial speed $v(0) \in [10, 90]$km/h. The system registers a collision when distance between two vehicles is lower than 2 m.

Applying the default LLM with maximum error of 5% on a 30% test set, we obtained 85,9% of accuracy, 75.4% sensitivity, 86.8% specificity and 46 % F1-score.

Reliability from Outside

From the value ranking for the collision class ($y = 1$), we obtained $PER >$0.43 and $F_0 \leq -7.50 \times 10^3$N as the first two most important intervals. We then applied the optimization approach as in Equation 5 and found $\delta^*_{s_1} = -0.034$, $\delta^*_{t_2} = -0.416$, with FNR=0 and TNR=0.34. Thus, according to the definition in Equation 6, the safety region we obtain is the following:

$$\mathcal{S}_1 = ((PER \in (0.2, 0.4154)) \wedge (F_0 \in (-4.37, -1) \times 10^3)$$

A visual representation of such region is in Fig. 3. Also, we tested the method with $N^{FR} = 3$, including the third most important interval from value ranking too, i.e. $N > 6$. We got $\delta^*_{s_1} = -0.184, \delta^*_{t_2} = -0.166$ and $\delta^*_{s_3} = -0.1$ with FNR=0 and TNR=0.19. In this case, the safety region is the following volume (Fig. 4):

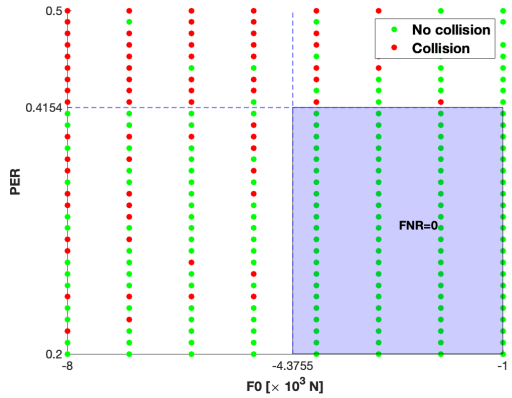

Figure 3: Scatter plot of the first two features (PER and F0) with representations of the safety region

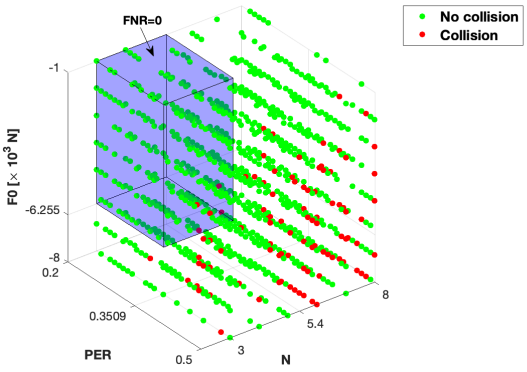

Figure 4: 3D scatter plot of the first three features (PER,F0,N): the safety region is represented by the volume (in violet)

$$\mathcal{V}_1 = ((PER \in (0.2, 0.3509)) \wedge (F_0 \in (-6.255, -1) \times 10^3) \wedge (N \in [3, 5])$$

Reliability from Inside

Following the optimization approach in Equation 7, we first chose the first two intervals from the value ranking of the safe class ($y = 0$): $PER \leq 0.33$ and $F_0 > -3.50 \times 10^3$N. Then, we computed the optimal threshold perturbations $\delta^*_{t_1} = 0.356, \delta^*_{s_2} = 0.686$, for which we got FNR=0 with TNR=0.13. The obtained safety region is the surface (Fig. 5):

$$\mathcal{S}_0 = (PER \in (0.2, 0.2125) \vee F_0 \in (-1.1001, -1) \times 10^3)$$

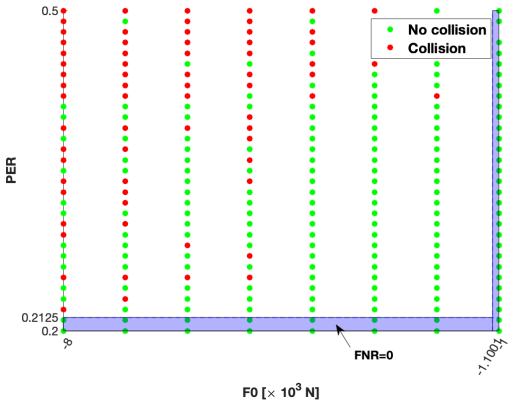

Figure 5: Scatter plot of the first two features (PER and F0) for safe class with representations of the safety region

Zero Error LLM

By lowering the LLM maximum error allowed to 0%, we were able to look for more complex safety regions. After training the LLM model with 0% error, we obtained 99 rules for the safe class and joined the first 4 with the highest covering. This corresponded to the following predictor:

IF ($N \leq 5$ **and** $v(0) \leq 54.50$) $\vee$

($PER \leq 0.295$ **and** $N \leq 7$ **and** $v(0) \leq 86.50$) $\vee$
($v(0) \leq 28.50$ **and** $PER \leq 0.445$) $\vee$
($v(0) \leq 28.50$ **and** $N \leq 6$ **and** $d(0) \leq 7.86$)

THEN ***safe***

Before any perturbation, the metrics are FNR=0.05 and TNR=0.55. We then exploited the feature ranking to individuate which features to tune for lowering FNR as much as possible. The two most influent features resulted were $v(0)$ and PER in this case. By solving Equation 9, we perturbed them: in this case, we were able to achieve only a suboptimal solution, with FNR=0.02 and TNR=0.45, corresponding to $\delta_{t1} = 0.000877$ for $v(0)$ and $\delta_{t2} = 0.277$ for PER. In order to maintain TNR as large as possible, we perturbed only the most stringent threshold where the same feature was present in more than one rule. The final result is a slight modification of rules above with respect to $v(0)$ (as the inherent δ is very low) and with a significant impact on PER as it is reduced to its lower bound in the second rule.

Discussion

Comparing the obtained results on the two datasets, we can see that inferring reliability from the available rules is highly dependent on the structure of the data under analysis. The three methods performed differently in the two cases. More specifically, the LLM 0% achieved optimality (zero FNR) for the fatigue problem and suboptimality (almost zero FNR) in platooning. However, as expected, in both cases results showed an improvement in TNR with respect to the other methods. On the other hand, the inside-outside methods show flexibility in looking at the feature space, alternating good results (outside in platooning in two dimensions), surprising results (outside in platooning in three dimensions is outperformed by the same in two dimensions) and bad results (inside in platooning in two dimensions). The outside approach finds larger (higher TNR) safety regions than the inside one both in fatigue and platooning. Inside-outside may be even joined together when the feature ranking agrees on the most important features for the available classes. As this happens in the platooning case, we may consider the safety regions involving PER and F_0 (Fig. 3 and Fig. 5), and, by visual analysis of the overlap of such regions (see Fig. 6), we could join them to find a larger and more complex safety region.

Conclusions and Future Works

In this work, we have studied native XAI models as a solution for safety insurance. In particular, we focused on a rule-based model, the LLM, and demonstrated how innovative rule optimization algorithms can be applied to design "safety regions" in the features space with zero statistical error. By testing and comparing our proposed methodologies on problem instances of different nature (physical fatigue and vehicle platooning), we have also shown how their performance varies between the datasets.

Future works may extend the testing through cross-validation in the presence of a large amount of data, including the adoption of data augmentation techniques. The characterization of the placement of the points deserves further study to understand the optimal covering of the safety regions. The translation of DL logic into rules with further design of safety envelope is another

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2022.3159098, IEEE Intelligent Systems

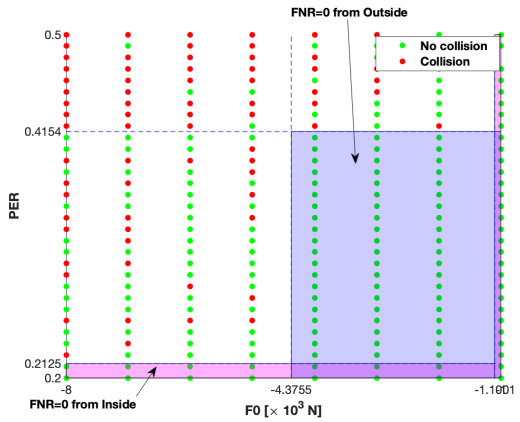

Figure 6: Scatter plot of the two most important features in vehicle platooning LLM classification (PER and F_0), with representation of the safety regions found with Inside (pink area) and Outside (blue area) methods: the overlap of such regions defines a new safety region, where TNR reaches higher values

topic we are going to pursue in the near future.

REFERENCES

1. R. Madhavan, J. A. Kerr, A. R. Corcos, and B. P. Isaacoff, "Toward trustworthy and responsible artificial intelligence policy development," *IEEE Intelligent Systems*, vol. 35, no. 05, pp. 103–108, sep 2020.
2. A. Clavière, E. Asselin, C. Garion, and C. Pagetti, "Safety Verification of Neural Network Controlled Systems," Nov. 2020, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02975455
3. "Standardization in the area of artificial intelligence," ISO/IEC, Washington, DC 20036, USA, Standard, Creation date 2017, https://www.iso.org/committee/6794475.html.
4. "Road vehicles safety of the intended functionality pd iso pas 21448:2019," International Organization for Standardization, Geneva, CH, Standard, Mar. 2019.
5. "Concepts of design assurance for neural networks codann," European Union Aviation Safety Angency, Daedalean, AG, Standard, Mar. 2020, also available as https://www.easa.europa.eu/sites/default/files/dfu/EASA-DDLN-Concepts-of-Design-Assurance-for-Neural-Networks-CoDANN.pdf.
6. V. N. Balasubramanian, S. Ho, and V. Vovk, *Conformal Prediction for Reliable Machine Learning*, 1st ed. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Elsevier, 2014.
7. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103
8. S. Mohseni, M. Pitale, V. Singh, and Z. Wang, "Practical solutions for machine learning safety in autonomous vehicles," *CoRR*, vol. abs/1912.09630, 2019. [Online]. Available: http://arxiv.org/abs/1912.09630
9. S. A. Seshia, A. Desai, T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, S. Shivakumar, M. Vazquez-Chanlatte, and X. Yue, "Formal specification for deep neural networks," in *Automated Technology for Verification and Analysis*, S. K. Lahiri and C. Wang, Eds. Cham: Springer International Publishing, 2018, pp. 20–34.
10. J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *CoRR*, vol. abs/1810.03292, 2018. [Online]. Available: http://arxiv.org/abs/1810.03292
11. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6405–6416.
12. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1321–1330. [Online]. Available: http://proceedings.mlr.press/v70/guo17a.html
13. X. Zhang, Z. Wang, D. Liu, and Q. Ling, "DADA: deep adversarial data augmentation for extremely low data regime classification," *CoRR*, vol. abs/1809.00981, 2018. [Online]. Available: http://arxiv.org/abs/1809.00981
14. D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019.
15. D. Isele, A. Nakhaei, and K. Fujimura, "Safe reinforcement learning on autonomous vehicles," *CoRR*, vol. abs/1910.00399, 2019. [Online]. Available: http://arxiv.org/abs/1910.00399
16. U. Becker, "Increasing safety of neural networks in

medical devices," in *Computer Safety, Reliability, and Security*, A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch, Eds. Cham: Springer International Publishing, 2019, pp. 127–136.

17. M. Muselli, "Switching neural networks: A new connectionist model for classification," pp. 23–30, 01 2005.
18. M. Mongelli, E. Ferrari, M. Muselli, and A. Fermi, "Performance validation of vehicle platooning through intelligible analytics," *IET Cyber-Physical Systems: Theory & Applications*, vol. 4, no. 2, pp. 120–127, 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cps.2018.5055
19. N. Williams, "The Borg Rating of Perceived Exertion (RPE) scale," *Occupational Medicine*, vol. 67, no. 5, pp. 404–405, 07 2017. [Online]. Available: https://doi.org/10.1093/occmed/kqx063
20. Z. Sedighi Maman, Y.-J. Chen, A. Baghdadi, S. Lombardo, L. A. Cavuoto, and F. M. Megahed, "A data analytic framework for physical fatigue management using wearable sensors," *Expert Systems with Applications*, vol. 155, p. 113405, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420302293

Sara Narteni Got her M.Sc. in Bioengineering at the University of Genoa on March 2020, with a thesis entitled "Pleural line ultrasound videos analysis for computer aided diagnosis in acute pulmonary failure". She is PhD student at Politecnico di Torino, working in the IEIIT institute of Consiglio Nazionale delle Ricerche. She works on data analytics and machine learning topics from different fields, such as industry, healthcare and automotive. Moreover, her research interests also concern computer security topics, including covert channels and Internet of Things.

Vanessa Orani Got a M.Sc. in Stochastics and Data Science in April 2019 with the thesis "Bayesian isotonic logistic regression via constrained splines: an application to estimate the serve advantage in professional tennis". Now she is research fellow at the laboratory IEIIT of the CNR, the main research activities concern machine learning applied in different field, including health, transport and telecommunications. She is currently involved in the ALISA project, funded by Regione Liguria and in cooperation with Aitek S.p.A. (www.aitek.it), to investigate AI approaches with data obtained from IoT devices and to deepen video content analysis/image processing.

Ivan Vaccari Got his Ph.D. in Computer Science and a Computer engineering degree cum laude at the University of Genoa, respectively in 2021 and 2017. During his research activities, he worked in different European projects focused on security in healthcare data, IoT and financial infrastructures. He is currently a research fellow at IEIIT institute of Consiglio Nazionale delle Ricerche, working on IoT and network security focused on identification of vulnerabilities and developed of innovative cyber threats. Regarding detection and mitigation systems, he is working on machine learning and artificial intelligence approaches.

Enrico Cambiaso Got his Ph.D. degree in Computer Science at the University of Genoa, while working for Ansaldo STS and Selex ES, both companies are part of the Finmeccanica group. He has a strong background as computer scientist and he is currently employed at the IEIIT institute of Consiglio Nazionale delle Ricerche, as a technologist working on cybersecurity topics and focusing on the design of last generation threats and related protection.

Maurizio Mongelli Obtained his Ph.D. Degree in Electronics and Computer Engineering from the University of Genoa (UNIGE) in 2004. The doctorate was funded by Selex Communications S.p.A. (Selex). He worked for both Selex and the Italian Telecommunications Consortium (CNIT) from 2001 until 2010. During his doctorate and in the following years, he worked on the quality of service for military networks with Selex. He was the CNIT technical coordinator of a research project concerning satellite emulation systems, funded by the European Space Agency; and he spent three months working on the project at the German Aerospace Center in Munich. He is now a researcher at the Institute of Electronics, Computer and Telecommunication Engineering (IEIIT) of the National Research Council (CNR), where he deals with machine learning applied to bioinformatics and cyber-physical systems. He is co-author of over 100 international scientific papers and 2 patents.