On the Intersection of Explainable and Reliable AI for physical fatigue prediction

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2022, accepted 4 July 2022, date of publication 18 July 2022, date of current version 25 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3191907

RESEARCH ARTICLE

On the Intersection of Explainable and Reliable AI for Physical Fatigue Prediction

SARA NARTENI[1,2], VANESSA ORANI[3], ENRICO CAMBIASO[1], MATTEO RUCCO[4], AND MAURIZIO MONGELLI[1], (Member, IEEE)

[1]Consiglio Nazionale delle Ricerche (CNR), IEIIT Institute, 16149 Genoa, Italy
[2]DAUIN Department, Politecnico di Torino, 10129 Turin, Italy
[3]Aitek S.p.A., 16122 Genoa, Italy
[4]Biocentis Ltd., 20144 Milan, Italy

Corresponding author: Enrico Cambiaso (enrico.cambiaso@ieiit.cnr.it)

This work was supported by the National Research Council (CNR)-Institute of Electronics, Information Engineering and Telecommunication (IEIIT) through Fondazione Compagnia di San Paolo, Scientific Research Call 2019 through the Bando (2019–2020) per Progetti di Ricerca Scientifica Presentati da enti Genovesi through Project "Advances in Pneumology via Information and Communication Technology (ICT) and Data Analytics" (PNEULYTICS) under Grant 2020.AAI22.U41/SD/pv, Grant 2019.0988, and Grant 34754.

ABSTRACT In the era of Industry 4.0, the use of Artificial Intelligence (AI) is widespread in occupational settings. Since dealing with human safety, explainability and trustworthiness of AI are even more important than achieving high accuracy. eXplainable AI (XAI) is investigated in this paper to detect physical fatigue during manual material handling task simulation. Besides comparing global rule-based XAI models (LLM and DT) to black-box models (NN, SVM, XGBoost) in terms of performance, we also compare global models with local ones (LIME over XGBoost). Surprisingly, global and local approaches achieve similar conclusions, in terms of feature importance. Moreover, an expansion from local rules to global rules is designed for Anchors, by posing an appropriate optimization method (Anchors coverage is enlarged from an original low value, 11%, up to 43%). As far as trustworthiness is concerned, rule sensitivity analysis drives the identification of optimized regions in the feature space, where physical fatigue is predicted with zero statistical error. The discovery of such "non-fatigue regions" helps certifying the organizational and clinical decision making.

INDEX TERMS Physical fatigue detection, industry 4.0, explainable AI, logic learning machine, LIME, anchors, reliable AI.

I. INTRODUCTION

A. HUMAN FATIGUE

1) CONTEXT

According to the Health Safety Executive, human fatigue is defined as *"a state of prolonged mental or physical exertion; it can affect people's performance and impair their mental alertness, which leads to dangerous errors"* [1]. If we consider the effects on a critical system, such general definition well suits to the industrial context.

The huge presence of automation in today's industrial environments, referred to as Industry 4.0, leads to an important discussion on how to properly integrate human workers with machines and technologies like Artificial Intelligence (AI) and expert systems [2]. Research in this field is now focusing on how to take advantage of machines and technology, without impairing human health during work. This is accomplished by studying the impact of physical tasks on human performance. In fact, it is known that a deteriorated physical condition can increase the probability of casualties, injuries, slips and falls during working time.

2) SAFETY

Traditionally, in industrial robotics, safety has been addressed through a rigid separation between human and robot operators; in Human-Robot Collaboration (HRC), which involves direct physical contact [3], several technologies for identifying and avoiding unintended contacts between humans and robots have been published. However, current formal methods for safety analysis of HRC do not recognize the

The associate editor coordinating the review of this manuscript and approving it for publication was Weipeng Jing.

operators as proactive factors. Hence, research is strongly focused on formal analysis of robots' behaviours, verified through model checking techniques. Particularly, although such techniques are suitable for checking the correctness of the robots' components, it is very difficult to have a complete formal deterministic description of the surrounding physical world. An alternative to model checking are simulation-based cyber-physical systems (CPSs) techniques, typically requiring models of the entire CPS, not always available in *collaborative robots* (cobots). Traditional formal approaches such as FMEA (Failure Mode and Effects Analysis) and FTA (Fault Tree Analysis) are not well-suited for HRC applications, as they do not capture hazards related to human factors or combinations of hazards. Another solution based on model checking and runtime verification of robots operating systems has been published [4]. Instead, a limited number of researches focus on data-driven approaches leveraging on formal methods and ISO standards [5].

3) STANDARDS

On the other hand, formal standards for industrial robotics are devoted to functional performance and safety, almost without considering any human safety requirement apart from physical ergonomic issues. Operators' psychological safety and ethical issues have not been highly addressed before, when robots in factories remained segregated away from human contact. Also, the need of advanced manufacturing training methods was quite limited, since humans were not supposed to directly collaborate with the robots. Industrial applications of HRC only started being addressed in international standards in the past decade, in some clauses of ISO 10218-1/-2.[1] Supplementing these requirements, ISO/TS 15066 was published in 2016 [6] to address the growing need for HRC standards. Such standards focus on technical safety and are not expected to venture in psychological and customized aspects.

4) INDUSTRY 4.0

With the advent of new technologies and the transition of production to Industry 4.0, a more flexible approach to manufacturing is pursued to achieve higher productivity, where robots and human operators are allowed to collaborate and interact.[2] Such transformation leads to overcoming traditional safety procedures and the development of new safety-assuring technologies for the minimization of risks connected with HRC. It is worth mentioning the concept of *safety agent* developed in the SHERLOCK project [3]: the approach is based on a predictive model of human motion, compared against the planned robot trajectory and online monitoring of satisfaction of safety requirements with formal methods.[4]

5) DEVICES

Many devices can be used to monitor human movements, such as electromiography (EMG), 3D optical tracking, infrared cameras or reflective markers, but the most useful ones are the Inertial Movement Units (IMUs). These sensors are able to measure acceleration, velocity and orientation in 3D space, by using a combination of accelerometers, gyroscopes and magnetometers [7], without being invasive for people.

B. MACHINE LEARNING

Machine learning (ML) and AI algorithms can come to help in analyzing data from human movement monitoring sensors, in order to distinguish fatigued and non-fatigued states. One of the main requirements for AI is the need of labelled data for classification. While there is lack of objective systems to quantify physical fatigue, many subjective methods exist, based on individual perception of fatigue, collected through questionnaires. The most common and widely used one is Borg RPE (Rate of Perceived Exertion) scale [8], which uses numbers from 6 to 20 for ranking fatigue levels from "None" to "Very, very hard". Other subjective scales exist: Borg CR10 scale, Need for Recovery Scale (NRS) and Fatigue Assessment Scale (FAS) [9].

Supervised ML methods like linear discriminant analysis, naive Bayes, artificial neural networks, k-nearest neighbors and, more often, support vector machines (SVM) are commonly adopted for human performance assessment, as they allow traditional classification tasks. Statistical models like logistic regression and penalized logistic regression may be useful to build a relationship between a discrete output and input variables (predictors). Furthermore, ensemble methods, i.e., a combination of single supervised classifiers, such as random forest, may be used for physical fatigue detection too [2]. Despite their huge development in recent years, deep learning models are not indicated in this field as they require very large datasets, which can't be available in this context.

C. XAI

In biomedical contexts, a qualitative information about the system is essential. The onset mechanisms of a pathology should be discussed with the medical staff. [10] For this reason, ML research is now oriented to a branch, commonly referred to as eXplainable Artificial Intelligence (XAI) [10], whose goal is providing understanding of the logic involved in ML-driven decisions. There are two main approaches to this field: one, also referred to as Interpretable ML [11], consists in learning transparent-by-design models that make predictions in an intelligible way, such as rule-based models; the other main group of techniques consists in finding explanations for black-box predictions (these methods are

[1] https://www.iso.org/standard/51330.html, https://www.iso.org/standard/41571.html

[2] https://sharework-project.eu/european-projects-on-human-robot-collaboration/

[3] https://www.sherlock-project.eu/home

[4] https://zenodo.org/record/3901416#.YCAG_OhKiUk

also known as post-hoc techniques) [12]. Another important classification of XAI techniques involves the distinction between global and local methods [13]: global methods aim at explaining the overall logic of a model, while local methods focus on explaining the reasons behind single decisions on single data instances [14].

The XAI topic then constitutes one of the main open challenges of the AI sector [15]–[17]. To the best of our knowledge, the use of XAI is still unexplored in the fatigue problem.

Combining both the above-mentioned categorizations of XAI techniques, our work involves two kinds of XAI techniques: (i) global transparent-by-design methods: Logic Learning Machine (LLM) and Decision Tree (DT); (ii)local post-hoc explanations of black-box models: Local Interpretable Model-agnostic Explanations (LIME) and Anchors.

D. RESPONSIBLE AI

Another advantage of XAI relies on driving *trustworthiness* of AI (also known as *Responsible AI*) [18]. We focus here on the specific sub-case of Responsible AI related to the intersection of AI with safety and call it *Reliable AI*. The attention of field experts is increasingly turning to AI impact on society: awareness is growing of the potential risk of serious accidents, for example due to design error, poorly specified goals or simple misapplication by the AI. The certification of an AI-driven system is therefore another open challenge of science and engineering for the near future (see, e.g., the recent family of standards dedicated to AI in the automotive sector [19]). The sensitivity analysis of XAI offers support to Reliable AI for the control [20], [21], formal validation [22], [23] and cybersecurity [23] of cyber-physical systems. The paper re-designs such sensitivity analysis for the fatigue problem in the safety perspective (i.e., finding explainable regions in which any operator is not fatigued for certain).

E. CONTRIBUTIONS

More specifically, this work is intended to provide complementary information to formal model-based solutions (i.e., safety agent) for maximizing human safety. The safety agent becomes active during the execution of safe actions and it supervises the operator's fatigue and informs the operators about the decisions that will be taken by the robot to minimize operator's fatigue.

The following novelties are presented in the paper, based on the XAI approaches mentioned in Sec. I-C:

- Full performance comparison of global transparent-by-design methods (LLM and DT) with black-box models (NN, SVM, XGBoost), in terms of accuracy, sensitivity, specificity and F1-score.
- Rule optimization to drive reliable AI (statistical zero error) of ''non-fatigue regions'' to facilitate the safety of workers on the field.
- Feature ranking-based comparison between global interpretable models (LLM and DT) and local post-hoc explanations of a black-box (XGBoost + LIME).
- Rule optimization to extend Anchors coverage of data.

Overall, the paper gives a complete vision of machine learning applied to the fatigue problem through black-box approaches and emphasis on native eXplainable (transparent-by-design) algorithms, as well as the further optimization of the latter for error control that impacts on safety through sensitivity analysis and variations of LLM and Anchors.

The remaining of the paper is structured as follows: in Section II we revised the most recent literature about AI solutions for physical fatigue detection. Instead, in Section III we describe the theoretical basis of the ML/AI methods we adopted in our work, while Section IV describes the adopted dataset, along with a presentation of all the phases of the work. In Section V we report and discuss the obtained results, while we conclude the paper and propose future works on the topic in Section VI.

II. RELATED WORK

In this section, we provide a literature review in the field of AI solutions for fatigue classification.

A. BLACK-BOX APPROACHES

Thanks to their practicality, Inertial Movement Unit (IMU) sensors are widely adopted for data collection in many studies. In [7], authors developed an integrated system for workers fatigue state recognition, called Smart Safety Helmet, i.e. a helmet in which IMU sensors are combined with electroencephalography (EEG) in order to detect physical exertion and mental stress at the same time. An AI module allows the detection and computes a threshold for fatigue. The device can generate a vibro-tactile feedback to the operator when the threshold is reached, and a wireless signal is sent to the industrial machines to stop their activity.

In [24], SVM with three different kernels (linear, polynomial and RBF) classifier was adopted to find differences in walking task before and after fatigue inducement; kinematic gait data were collected from a combination of IMUs and passive infra-red markers put on lower limbs.

Instead, the authors of [25] developed a data-driven approach for whole-body fatigue detection in manufacturing environment through the use of inexpensive wearable IMUs collocated on different body parts and a heart rate monitor device. LASSO model has been chosen as the best one for detection, compared to other penalized logistic regression models e.g. elastic-net and ridge regression, as it was the optimal combination of amount of variation explained, performance and applicability (with the lowest number of features). As data were unbalanced between fatigued and non-fatigued, Random Under Sampling (RUS) technique was applied. A set of predominant features was derived by ranking model's coefficients. The results highlighted a predominance of wrist movements in determining fatigue.

Following the previously cited work, [26] studied a method to distinguish fatigued from non-fatigued subjects using data from a single IMU on the right ankle. After a raw data pre-processing phase to compute gait kinematics measurements and derive so-called motor templates, a Euclidean distance-based algorithm for template matching ($1 Recognizer [27]) was applied to extract distance scores. Along with step duration, such scores are used as features for RBF SVM classification.

Authors in [28] adopted jerk, the time derivative of acceleration, as a measure of physical fatigue in a bricklaying task. Data were collected by an IMU-based motion capture suit and a SVM model was applied to recognize intra-subject and inter-subject fatigue. Results showed a better performance for the intra-subject case. In [29], a random forest classification with Bayesian inference model was adopted to detect different modes of load carriage tasks with different load levels, which is known to alter gait patterns and pelvic-thoracic coordination. For this reason, inertial sensors were put both on thorax and shanks. From the derived predictive model, an interpretation was provided for each predictor variable by computing the Gini impurity index.

In [30], a computationally efficient method based on acceleration profile from a single IMU put on the ankle was developed. It consisted in a Statistical Process Control technique, where deviations from a baseline, non-fatigued, dataset were monitored.

While most of the studies involved data acquisition in laboratory settings, [31] proposed a method for fatigue recognition in walking task directly in the workplace; it is based on smartphone-integrated IMU data. SVM with RBF kernel algorithm is applied for 2-class (fatigued/not-fatigued state) and a more useful 4-class (no fatigue, low, medium and high fatigue) detection.

IMU sensors are not the only devices used in literature for fatigue classification. In [32], 3D optical tracking was exploited for exhaustion detection in natural walking. In this analysis, linear discriminant analysis, naive Bayes, k-nearest neighbors and SVM performances were compared. In [33], flexible textile strain sensors and a random forest ML model are chosen to detect fatigue level in runners. Surface electromiography (sEMG) is instead adopted in [34] in order to detect neck muscle fatigue while using mobile phones; logistic regression, decision tree and SVM classifiers were compared for this purpose. SVM resulted as the best one.

Ensemble tree learning methods, like Random Forests, consist in aggregating multiple decision trees to improve the generalization capabilities and avoiding overfitting. Nevertheless, the aggregation of several models makes it more complex to interpret compared to individual tree models. Similarly, SVM, despite being widely adopted for the problem of physical fatigue modelling, is even more complex and less understandable, especially with the increase of feature space dimensions. Hence, we decided to move the analysis towards XAI-based solutions.

B. THE XAI AND ERROR ANALYSIS OPEN ISSUE

Our work involves both main categories of XAI methods, i.e. interpretable methods and black-box explainers (see Sec.I-C).

Referring to interpretable models, we focus on rule-based ones. Many different algorithms have been proposed in this field of research, dealing with either rule sets of unordered rules or rule lists with ordered rules [17], [35]. In [36], a Bayesian framework is introduced to learn a low number of short rules, seeking a balance between accuracy and interpretability through user-adjustable Bayesian prior parameters. [37] proposes EXPLORE algorithm to induce disjunctive decision rules (rule sets) in a systematic and efficient manner, based on a branch-and-bound approach that relies on user-defined performance constraints and seeks rule optimal length. In [38] generalized linear models using rule-based features are considered for regression and probabilistic classification; this solution trades off rule set complexity, in terms of number and length of rules, and prediction accuracy. Multi-value Rule Set (MRS) provides a more concise and feature-efficient model form for classification and explanation [39]. Always seeking a balance between accuracy and interpretability, authors in [40] study interpretable decision sets, i.e. sets of independent if-then rules.The learning process is based on an objective function that optimizes accuracy and interpretability of the rules using smooth local search. Rule lists are also widely investigated, being collections of ordered rules [35]. Falling rule lists [41] have been introduced, providing both predictive modelling and a descending ranking of rule outcome probability (i.e., rules leading to higher class probabilities are reported first). [42] presents an innovative technique for rule lists generation over categorical feature space, called CORELS, that gives the optimal solution according to the training objective, along with a certificate of optimality. Also, in [43] Bayesian rule lists method is investigated, combining decision lists with Bayes approaches, with applications in healthcare contexts. Rule-based models have been studied in hybrid approaches too, where they are combined to black-box models and can substitute the black-box decisions for given subsets of data [44]. Companion rule lists [45] are based on a rule list combined with a pre-trained black-box model, where users are allowed to switch between rules and the black-box, based on their requirements for interpretability or accuracy.

In the context of causal inference, [46] introduces an innovative method for learning causal rule sets, based on simulated annealing. In this work, a crucial point of rule-based models is stressed out: according to the complexity (noise) of available classes of data, the number and size of the generated rules can vary a lot (i.e., complex boundaries may lead to high numbers of small rules and viceversa). This aspect will be also outlined in our rule-based approach to the design of non-fatigue regions (Sections IV-D and V-B).

Another set of approaches to XAI is black-box (or post-hoc) explanation, which can be found as model explanations, outcome explanations or model inspections [17]. Model

explanations refer to global interpretable models used to learn the predictions of a black-box [47], [48], also involving rule extraction [49]. Outcome explanations methods consist in making single instances predictions interpretable: examples are LIME and Anchors, that will be used in this paper. Model inspections consist in understanding specific properties of the black-box, for example via providing tree-like visual interpretations of the black-box decisions [50] or studying the effects of changing attributes via sensitivity analysis [51], [52]. Local black-box outcome explanations find relevant applications for Deep Neural Networks, where several data-driven and model-driven techniques have been developed [53], [54].

In the specific context of physical fatigue detection, to the best of our knowledge, only few studies can be found adopting any XAI technique and these are mainly based on feature importance [55], [56]. A rule-based (IF-THEN) fuzzy classifier has been used in [57] for classifying local muscle fatigue into three states (non-fatigued, transition-to-fatigue, fatigued). However, in this case data were collected from sEMG signals, not IMU data, and fatigue was induced by asking participants to maintain a 90° elbow angle.

In our paper, we move a step further by comparing global rule-based interpretable models with local post-hoc explanations methods. Moreover, we attempt to discover regions where false negatives tend to zero through rule-based models, which is new and in line with Responsible AI paradigm.

In this perspective, the open issues discussed by [58] (in the error analysis and critique sections) are coherent with the present idea to calibrate the value ranking through a controllable margin from the fatigued to the non-fatigued (safe) class. More specifically, [58] discusses the tuning of false positives and false negatives through class weights during training. Such methodology opens the door to more sophisticated approaches, such as through the exploration of zero error surfaces.

III. PRELIMINARIES ON SUPERVISED ALGORITHMS

In our work, we compared the classification performances obtained through different ML methods, including black-box, and analyzed the behaviour of global interpretable methods (LLM, DT) with respect to local methods for the explanation of black-box models (LIME [59], Anchors [60]). Moreover, we addressed particular attention to the tuning of LLM rules conditions to individuate "non-fatigued regions" (see IV-D), by exploiting its feature and value rankings. In this section, an overall description of such methods' fundamentals is then provided. It is divided into the following sections: Section III-A for global transparent-by-design models, Section III-B for local post-hoc explanations, Section III-C for black-box models. Figure 1 provides a general overview of supervised learning as categorized throughout this Section.

A. GLOBAL TRANSPARENT-BY-DESIGN MODELS

Global transparent-by-design models provide predictive models in an intelligible way. In this work, we adopted

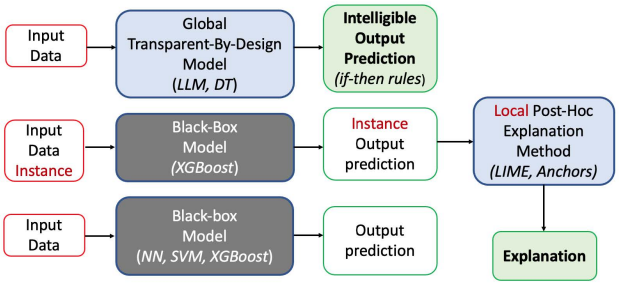

FIGURE 1. High-level description of the categories of supervised learning methods adopted in this paper. At the top, global transparent-by-design models provide predictions based on the input data in an interpretable way (if-then rules in our case). In the middle, local post-hoc explanations of black-box models provide explanations for single black-box predictions on single data instances. At the bottom, black-box models provide predictions, based on the data, without any kind of explanation.

two rule-based models, the LLM (Sec. III-A1) and the DT (Sec.III-A2), both coming as human-understandable *if-then* rules. The fundamentals of the methods are here presented, along with the associated classification scoring system (Sec. III-A3) and their properties of feature and value ranking (Sec. III-A4).

1) LOGIC LEARNING MACHINE (LLM)

Logic Learning Machine is an innovative supervised method; it was developed as an efficient implementation of Switching Neural Networks [61]. LLM has the aim of building a classifier $g(x)$ described by a set of rules structured as follows: **if** $<$*premise*$>$ **then** $<$*consequence*$>$. The $<$*premise*$>$ is a logical product (AND) of conditions on the input features, whereas $<$*consequence*$>$ corresponds to the output class. The generation of the model is a three-step process:

1) *Discretization and mapping to a Boolean lattice (latticization, [62])*: each variable is transformed into a string of binary data in a proper Boolean lattice, using the inverse only-one code binarization. All the strings are eventually concatenated in one unique large string per each sample.
2) *Shadow Clustering*: a set of binary values, called implicants, are generated, which allow the identification of groups of points associated with a specific class.
3) *Rule generation*: all the implicants are transformed into a collection of simple conditions and eventually combined into a set of intelligible rules.

Generally speaking, for any rule-based model, the building of the conditions in a rule is done by jointly considering all the involved variables: hence, the conditions in a rule are always dependent. For the LLM, the conditions are built based on the implicants, which are defined as binary strings in a Boolean lattice that uniquely determine a group of points associated with a given class. Such clusters of points (in the Boolean space) are then translated into rules that combine a subset of joint variables. It is straightforward to derive from an implicant an intelligible rule having in its premise a logical product of threshold 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; in this way, resulting rules can overlap and represent different relevant aspects of the underlying phenomenon [63], [64].

A practical example on the LLM model building can be found in the Supplementary Materials of [62].

2) DECISION TREE (DT)

Decision trees are tree-based classifiers, based on the *divide-and-conquer* paradigm, that is partitioning the feature space in an iterative way, by selecting the best feature to split the data according to some statistical metric, such as information gain, gain ratio, Gini index or misclassification error. The structure of a DT is a graph where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision nodes are used to make a decision and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches. By navigating from leaf to root, it's possible to identify simple intelligible rules [64].

In DT, the choice of a variable, used to separate the classes through a proper threshold, is done node-by-node. Each step (node) of the tree construction depends on the choices (variables and thresholds) made previously, during the building of the upper part of the tree (till to that node).

The *divide-and-conquer* approach subsequently adds conditions to the tree based on smaller and smaller subsets of training data, thus lowering the computational time. On the other hand, it has the side effect of reducing the information available when selecting conditions after the first one.

As a consequence, the rules in the resulting models are disjoint. For this reason, DT may be over-sensitive to highly relevant attributes and imbalanced datasets [65], may depend too much on the training set or may be corrupted by irrelevant attributes and noise [20].

3) RULE-BASED CLASSIFICATION SCORING

We now introduce some general notation for any rule-based model (hence, valid for either LLM or DT).

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:

$$X_j > s, \quad X_j \leq t, \quad s < X_j \leq t, \tag{1}$$

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

For each rule generated by the algorithm, it is possible to define a confusion matrix associated to the rule. 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 useful metrics:

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

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

The covering $C(\mathbf{r}_k)$ is adopted as a measure of relevance for a rule $\mathbf{r}_k$; as a matter of fact, 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 classified by the rule.

Covering and error are both useful to determine the classification scores that are used to assign a class to input data [66].

Let $H_{\hat{y}}$ be the set of rules r_k predicting class $\hat{y}$ and satisfied by an input sample x_j. A score for every class is then derived as:

$$w_{\hat{y}} = 1 - \prod_{r_k \in H_{\hat{y}}} (1 - C(r_k))(1 - E(r_k)) \tag{4}$$

and every input x_j is assigned to the class with the highest score.

4) FEATURE AND VALUE RANKING

Feature and value rankings represent very useful methods to inspect the results obtained through rule-based models, like LLM or DT. Again, covering and error provide the basis for their definitions.

Feature ranking (FR) provides a way to rank the features included into the rules according to a measure of relevance. In order to obtain such measure of relevance $R(c_{l_k})$ for a condition c_{l_k}, we consider rule $\mathbf{r}_k$ in which the condition occurs, and the same rule without that condition, 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 refers to a specific variable X_j and is verified by some values $v_j \in \mathcal{X}$. In this way, a measure of relevance $R_{\hat{y}}(v_j)$ for every value assumed by X_j is derived by the following equation 5:

$$R_{\hat{y}}(v_j) = 1 - \prod_k \left(1 - R\left(c_{l_k}\right)\right) \tag{5}$$

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

B. LOCAL POST-HOC EXPLANATIONS OF BLACK-BOX MODELS

Along with models that are explainable-by-design, algorithms aiming to find explanations for black-models are gaining relevance in the last few years. In this work, we chose

LIME [59] and its improvement Anchors [60] as for comparison with LLM and DT. In the following Section III-B1 and Section III-B2, we describe how these algorithms work.

1) LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

Local Interpretable Model-agnostic Explanations, more often known as LIME, is an algorithm that can explain the results of any classifier by approximating it *locally* with an interpretable model [59]. It is a concrete implementation of local surrogate models, which are interpretable models used to explain individual predictions of black box machine learning models [67]. LIME is built upon three criteria that an explainer should follow [59]: being *interpretable*, so that one can understand a qualitative inputs-output relationship; being *locally faithful*, i.e. the explanations must correspond to how the model behaves in the proximity of the instance being predicted (without need of being good global approximations); finally, the explainer should be *model-agnostic*, that means being able to explain any kind of model.

On the basis of these criteria, LIME model is developed.

Consider $f(x)$ any black-box model making predictions for data instance x; also, let g be an interpretable model among all the available ones (e.g., linear models) and let π_x be a measure of the locality around x (it depends on a given distance function). LIME explanations are obtained by solving the following optimization:

$$\arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{6}$$

where function $\mathcal{L}$ defines the local fidelity between f and g, based on the locality defined by π_x. Function $\Omega(g)$ expresses the complexity of the interpretable model g. Being LIME a model-agnostic algorithm, no assumption is made on f, but it is only used to make predictions over new samples z obtained after adequately perturbing the instance x. The problem in Equation III-B1 is then solved on the perturbed dataset [59], [67].

2) ANCHORS

Anchors is a local-explanations method that has been developed from the same authors of LIME [60], in order to overcome the lack of accuracy of LIME explanations when moving away from the explained instance. In fact, *anchors* are high-precision rules, so that in the region of feature space where they hold, the same predictions are almost always guaranteed even on unseen instances. They're structured as if-then rules, so are intuitive, easy to understand and, above all, they allow a clear definition of coverage, intended as a measure of how many unseen instances they apply to [67]. Like LIME, Anchors is model-agnostic: it works by perturbing instances of interest according to a perturbation distribution D_x that must use an interpretable representation of the input.

Formally, a rule A acting on such an interpretable representation is an anchor if $A(x) = 1$ (i.e., all the rule predicates are true for x) and A is a sufficient condition for $f(x)$ with high probability, being $f(x)$ the black box model to explain. An anchor is therefore defined as follows:

$$\mathbb{E}_{D_x(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau, \quad A(x) = 1 \tag{7}$$

In equation 7, x is the instance to explain, A is the anchor, f is the black box model, $D_x(z|A)$ indicates the distribution of neighbors of x when the anchor applies, $0 \leq \tau \leq 1$ is a precision threshold.

Given an instance x, an anchor A has to be found, such that it applies to x, while the same class predicted for x gets predicted for at least τ of x's neighbors. A rule's precision results from evaluating neighbors or perturbations using the indicator function $\mathbb{1}_{f(x)=f(z)}$. Even if Anchors uses the same perturbations as LIME, they are faithful by design and adapt their coverage to the model's behaviour, so that the conditions in the anchors individuate wider regions when the underlying black-box decision function is wider too.

C. BLACK-BOX MODELS

In the context of supervised learning, black-box algorithms execute learning tasks without providing the users with sufficient information on *why* and *how* related results are obtained. This is due to their complexity, which is not enough human-friendly. Even if they can be more accurate than simpler interpretable models such as linear or logistic regression, they're not adequate when dealing with critical fields such as health and human wellness, like in our physical fatigue assessment case, where understanding algorithms outcomes is fundamental.

In the following sections, we summarize the foundations of two black-box methods, namely neural networks and support vector machine.

1) NEURAL NETWORKS (NN)

Artificial Neural Networks are biologically-inspired models formed by the interconnection of simple units, called neurons, arranged in layers. Each neuron performs a weighted sum of its inputs (generated by the previous layer) and applies a proper activation function to obtain the output value that will be propagated to the following layer. The first layer of neurons is fed by the components of the input vector $\mathbf{x}$, whereas the last layer produces the output class to be assigned to $\mathbf{x}$. The layers between input and output are called *hidden layers* [63]. Training a NN means finding the optimal weights for inputs of each layer that minimize a cost function and this is achieved by backpropagation. It consists of two parts: forward phase, in which, starting from the input vector, the output of the network is produced; the backward phase, where a gradient descent technique is applied to identify the correction of the weights to be implemented.

2) SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines are a a non-probabilistic binary linear classifier based on the idea of finding the hyperplane that optimally separates the data of two classes. Given a training set of labelled data, the algorithm finds the optimal

parameters of an hyperplane so to maximize the distance between the closest points of the two classes. These points are called *support vectors*.

The SVM classifier is linear, but it's possible to use *kernel functions*, i.e. non-linear transformations which convert the original feature space to a new higher dimensional feature space in which the linear classifier can be inferred. Most common kernels are polynomial, Gaussian, Radial Basis Function and sigmoid.

3) XGBoost

XGBoost stands for eXtreme Gradient Boosting and is now one of the most used ensemble algorithms in ML. It was developed by [68] as an optimization of Gradient Boosting, which guarantees higher performances. XGBoost is a tree ensemble model, that is a set of classification and regression trees (CART) used as base learners. Since one tree might not be enough to obtain good results, multiple CARTs can be used together and the final prediction is the sum of each CART's score. While in the original Gradient Boosting model the trees are built in series, XGBoost does it in a parallel way, resulting in a greater computational speed [69].

IV. MATERIALS AND METHODS

In this section, the adopted dataset, the experimental task and the accomplished ML tests are described in detail.

A. DATASET AND EXPERIMENTAL TASK DESCRIPTION

The data used in our study belong to an open-source dataset,[5] that was created by the authors of [2], who developed a data analytics framework for physical fatigue management made up of four steps (detection of fatigued state, identification of key features, diagnosis and recovery).

15 participants were asked to perform Manual Material Handling task, as a simulation of an industrial task, for 180 minutes and provide their RPE using the Borg Scale [8] every 10 minutes. A RPE$\geq$13 corresponds to a fatigued state, whereas lower values constitute the non-fatigued class. The executed task simulated warehousing operations by picking cartons (whose weight was 10kg, 18kg or 26kg), loading them on a 2-wheeled dolly, transporting them to a destination, and then palletizing them at the destination in a known order [25].

While performing the task, four IMUs (placed at the ankle, hip, wrist and chest of the participants) collected acceleration data, from which jerk and 3D space posture measures were derived. Furthermore, a heart rate monitor device was put on the chest, and the %HRR (Heart Rate Reserve) was calculated to express the percentage of an individual's heart rate being used under effort. For all these kinds of measure, mean and coefficient of variation (CV) were computed. In addition, individual (age, gender) and biomechanical features were considered. For a complete list of the original features, see Table 2 in [2], whereas in Table 1 we report the features adopted by us.

[5]https://github.com/zahrame/FatigueManagement.github.io/tree/master/Data

TABLE 1. List of the adopted features.

Feature
Age
Wrist Jerk Mean
Wrist Acceleration Mean
Hip Jerk Mean
Hip Acceleration Mean
Hip x posture Mean
Hip y posture Mean
Hip z posture Mean
Chest Jerk Mean
Chest Acceleration Mean
Chest x posture Mean
Chest y posture Mean
Chest z posture Mean
Ankle jerk Mean
Ankle Acceleration Mean
Ankle x posture Mean
Number of steps
Average step time
Average step distance
Time bent
Average back bent angle
Mean hip oscillation
Mean foot oscillation
Leg rotational velocity sag plane
Leg rotational position sag plane
Average vertical impact
Back rotation position in sag plane
Wrist jerk coefficient of variation
Wrist Acceleration coefficient of variation
Hip jerk coefficient of variation
Hip Acceleration coefficient of variation
Chest jerk coefficient of variation
Chest Acceleration coefficient of variation
Ankle jerk coefficient of variation
Ankle Acceleration coefficient of variation
Hip y posture coefficient of variation
Chest y posture coefficient of variation
Ankle y posture coefficient of variation

B. DATA PRE-PROCESSING

Starting from the previously introduced dataset, we decided to remove gender from the dataset since we chose to investigate the subjects' fatigue stratification - as will be shown later for the non-fatigue regions - based on their age only (under 40 and over 40). This choice is also supported by the results obtained in the original study that provided the data [2], where the feature selection process after the application of ML models individuated an impact of the age but not of the gender (for more information see Sec. 4.1.4 in the mentioned reference [2]). Moreover, up to now the sensitivity methods adopted for reliable fatigue detection via LLM (Sec. IV-D) are designed for numerical variables and not yet for categorical. However, we want to point out that, provided to extend those methods to categorical variables, the role of gender attribute could be investigated in the same way as we did for the age attribute. Further elaborations of the methods to include categorical attributes will be object of future studies.

Furthermore, in this work, we decided to focus on just one kind of sensors, i.e., the Inertial Movement Units (IMU) to monitor the accelerations associated to people's motion in the 3D space (then elaborated to extract features by the dataset creators [2]). IMU data are currently widely used to quantify physical fatigue through black-box approaches [24] and or goal is going a step further by investigating XAI role. For this reason, heart rate data were removed from the original dataset. A list of the adopted features is reported in Table 1

After this preliminary cleanage, we normalized data so that their scale was uniform, by applying z-score transformation. This step was carried in Matlab R2019a.

C. FATIGUE CLASSIFICATION

Once having completed the pre-processing of the data, we proceeded with data classification for the fatigue detection purpose. The major part of the analysis was conducted in Rulex software platform, a user-friendly data analytics platform, developed and distributed by Rulex Inc (https://www.rulex.ai), which allows to conduct entire machine learning tasks through drag-and-drop interface. Since it's not present in the functionalities offered by Rulex platform, we also used Python 3.8 *lime* package [59] for LIME algorithm and *anchor* for Anchors method [60], along with the underlying black-box method (XGBoost).

At the beginning of our work, we performed a comparison between XAI rule-based methods and black-box methods.

We first applied logic learning machine and decision tree. The dataset was imported in Rulex and split into a 67% of training set and 33% of test set with a fixed seed in order to maintain the reproducibility of our results. LLM and DT models, with the default settings provided by Rulex, were then built and trained. For LLM, default settings include the minimization of the number of conditions in the rules and a 5% of maximum error allowed for each rule.

Then, we repeated the classification tasks with LLM and DT by considering only the first three features obtained from the previous test's feature rankings.

Since LLM-based rules, in contrast with DT-based rules, can overlap, a higher precision is usually offered by such model. For this reason, we decided to focus more on LLM than DT and repeat the LLM classification by trying different combinations of the configuration parameters. In details, we changed the default LLM parameters as follows: the minimization of number of conditions (true by default) was removed, keeping the default maximum error rate; then, we changed the error rate (lowering it to 2% in one case and raising it to 10%) while leaving the minimal number of conditions; finally, we tried with a maximum error of 10% without minimization of the number of conditions. For comparison with black-box methods, we performed the classification task using a Rulex neural network model. Again, the dataset was split and the NN was trained. Different configurations of parameters (number of layers, number of neurons per hidden layer, learning rate) were tried to find the best results.

In addition, for sake of completeness, we also applied SVM method, as it is probably the most frequently used ML algorithm in the field of fatigue detection.

D. RELIABLE FATIGUE DETECTION

The problem of physical fatigue detection in occupational tasks is important from a clinical point of view. Identifying it with the lowest error possible is fundamental, allowing clinicians to make appropriate recovery interventions. In this context, intelligible models outperform other types of algorithms, since we can force the model to define ''non-fatigue regions'', in which the number of false negatives tends to zero. To this aim, the focus is therefore on finding envelopes ensuring a non-fatigued status. From a practical point of view, this can be thought as a safety guarantee: as soon as the subject parameters move outside of the individuated ''non-fatigue regions'' limits, an alarm might be generated to advice subjects to stop working and recover, or to inform the management about their deterioration.

The proposed design of such regions rely on the LLM model and its *feature and value ranking* properties [20].

Let X be a $D \times N$ matrix of all the input vectors $x_i \in \mathbb{R}^N$, with the total number of features N and $i \in [1, D]$. Let $g(x_i) = y$ be the function describing the LLM classification. In our case, $g(x_i) = 1$ if the prediction of the subject is fatigued and zero otherwise ($g(x_i) = 0$). Let D_1 be the number of fatigued instances and D_0 the number of non fatigued instances, so that $D_1 + D_0 = D$.

1) SENSITIVITY FROM OUTSIDE

Let N^{FR} be the number of the most significant features obtained through the feature ranking for the fatigued 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 the fatigued class as $[s_j, t_j]$. In our method, we expand such intervals 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 the non-fatigued class, namely, it finds the most suitable shape, in terms of rule-based intervals, of fatigue points around the non-fatigue ones. It is as follows:

$$\Delta^* = arg \min_{\Delta: N_1 = D_1} \mathcal{V}(\mathcal{P}(\Delta)) \tag{8}$$

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

$$\mathcal{S}_1 = ((-\infty, s_1 - \delta^*_{s_1} \cdot s_1) \vee (t_1 + \delta^*_{t_1} \cdot t_1, \infty)) \wedge ((-\infty, s_2 - \delta^*_{s_2} \cdot s_2) \vee (t_2 + \delta^*_{t_2} \cdot t_2, \infty)) \quad (9)$$

2) SENSITIVITY FROM INSIDE

An alternative way to perform the same search for "non-fatigue regions" consists in considering the N^{FR} most important features for non-fatigued ($y = 0$) class and reducing their most relevant intervals (again, provided by LLM value ranking) until the obtained region only contains true negative instances.

In this case, with the same notation as for the previous definition (section IV-D1), 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 8 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 non-fatigue class, until a fatigued point is reached. It is as follows:

$$\Delta^* = arg \max_{\Delta: N_1 = 0} \mathcal{V}(\mathcal{P}_0(\Delta)) \quad (10)$$

For $N^{FR} = 2$, the "non-fatigue region" is the following rectangle $\mathcal{S}_0$:

$$\mathcal{S}_0 = (s_1 + \delta^*_{s_1} \cdot s_1, t_1 - \delta^*_{t_1} \cdot t_1) \vee (s_2 + \delta^*_{s_2} \cdot s_2, t_2 - \delta^*_{t_2} \cdot t_2) \quad (11)$$

The results section V-B shows several examples of those optimal rectangles. The approach further helps discriminate further feature stratifications, e.g., by focusing on the age of the subjects lying on those regions. This stimulates the further inspection by the experts in the field (e.g., clinicians). Some stratification examples are included in the results.

3) 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 statistical error constraint and by starting from the available rule baseline. Zero error classification (for the non-fatigued class) is readily available by the shadow clustering adopted by LLM. The clustering process is applied with the further constraint of building clusters without superposition of points of more than one class [20] (LLM 0%, in the following). All the resulting rules with zero error are then joined in logical OR ($\vee$), thus describing a more complex geometry 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 procedure provides a set of m^0 rules $\mathbf{r}^0_k$, $k = 1, \ldots, m^0$ for the non-fatigued class ($y = 0$). Also, let $c^0_{l_k}$, $l^0_k = (1, \ldots, d^0_k)$ be the set of d^0_k conditions inside of each rule $\mathbf{r}^0_k$. We can join all the obtained rules $\mathbf{r}^0_k$ in logical OR operation ($\vee$), thus 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 $c^0_{l_k}$, 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 perturbation 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:

$$\boldsymbol{\delta}^* = arg \max_{\boldsymbol{\delta}: E(\hat{r}(\boldsymbol{\delta})) = 0} C(\hat{r}(\boldsymbol{\delta})) \quad (12)$$

E. GLOBAL AND LOCAL EXPLANATIONS COMPARISON

Besides the applicative context of physical fatigue detection, another important goal of our work is to compare global interpretable methods (LLM, DT) with local explanations of black box methods obtained through LIME and Anchors.

In order to perform such a comparison, LIME and Anchors algorithms were applied to the predictions of an XGBoost classifier. The training/test proportion was maintained of 67%/33% as for the previous Rulex analysis (Section IV-C).

As LIME and Anchors allow *local* explanations for single observations only, they were applied to all the test set records separately. For each instance, LIME provided a set of explanations with the corresponding weight. Anchors provided an if-then rule with related coverage measure instead.

We adopted the submodular pick LIME version (SP-LIME) of the algorithm in order to select a set of B most representative LIME explanations - derived from B data instances - that covered, together, the maximum number of features in a non-redundant way, thus giving a global understanding of the model. The parameter B is defined by the authors of [59] as the number of explanations needed by the human user to understand a model. In our case, we set $B = 5$.

To perform the comparison of LIME and Anchors with LLM and DT, we followed different approaches.

As regards LIME, we exploited the definition of importance of an explanation given in [59] to infer a feature importance metric.

We built a matrix F of size $n \times d$, where n is the number of explanations (equal to the number of instances in the test set in our case), d is the number of features involved in the classification problem and the element F_{ij} is the weight of feature j involved in the *i-th* LIME explanation, as found by LIME itself. We then computed the importance measure for each feature as in equation 13:

$$I_j = \sqrt{\sum_{i=1}^{N} |F_{ij}|} \quad (13)$$

In the case of Anchors, we compared its local if-then rules with the rules obtained with LLM and DT. Since a covering

measure is easily derived by the *anchor* Python library [60], we adopted it to rank the generated rules.

Thinking about the different nature of the three methods (i.e., global for LLM and DT, local for Anchors), the value of covering was expected to be quite different, being lower for Anchors.

After having verified our hypothesis by comparing the covering values for LLM/DT and Anchors, we assessed if it was possible to expand the Anchors coverage to include as many test set points as possible. To do so, we adopted an optimization approach based on rules conditions tuning.

Suppose to consider m Anchor rules $a_k, k = 1, \ldots, m$, each containing d_k conditions $c_{l_k}, l_k = 1_k, \ldots, d_k$. We define a perturbation vector $\boldsymbol{\delta}$ of elements $\delta_{l_k}, l_k = 1_k, \ldots, d_k$ acting on condition c_{l_k}. In this way, each Anchor can be modified on the basis of $\boldsymbol{\delta}$, so rule a_k can be expressed as a new rule $a_k(\boldsymbol{\delta})$, for which covering $C(a_k(\boldsymbol{\delta}))$ and error $E(a_k(\boldsymbol{\delta}))$ metrics can be computed just like in equations 2 and 3.

Our goal is then to find the optimal $\boldsymbol{\delta}^*$ as follows:

$$\boldsymbol{\delta}^* = arg \max_{\boldsymbol{\delta}: E(a_k(\boldsymbol{\delta})) < 0.05} C(a_k(\boldsymbol{\delta})) \tag{14}$$

V. RESULTS AND DISCUSSION

In this section, we provide an extensive performance evaluation for the different proposed approaches. In Section V-A, the results of the performance comparison between global rule-based and black box models are presented and, in Section V-B, the obtained "non-fatigue regions" designed in Section IV-D are shown. In Section V-C, SP-LIME results are reported and discussed. In Section V-D, LLM, DT and LIME feature rankings are illustrated and compared. Finally,in Section V-E we present our comparison between LLM, DT and Anchors.

A. FATIGUE DETECTION PERFORMANCE

In order to evaluate the performance of each ML algorithm we have tested, we adopted some common metrics derived from the *confusion matrix*. Being TP, TN, FP, FN the true positives, true negative, false positives and false negatives respectively, the following quantities can be computed:

- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$: it is a measure of the overall ability of the model of giving correct predictions.
- $Sensitivity = \frac{TP}{TP+FN}$: also known as True Positive Rate (TPR) it is the ability of the predictor in detecting fatigued state (class 1).
- $Specificity = \frac{TN}{FP+TN}$: also known as True Negative Rate (TNR), it measures the correct classification of non fatigued cases (class 0).
- $F1\text{-}score = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$: it is the harmonic mean of precision and sensitivity; it can vary in [0,1], with 1 being the F1-score for an ideal classifier.

In Table 2 the results of our comparison between rule-based XAI methods (LLM and DT) and black-box methods (NN, SVM and XGBoost) are shown.

TABLE 2. Performance measures of explainable rule-based methods vs black-box methods.

Method	Accuracy	Sensitivity	Specificity	F1-score
LLM	0.82	0.71	0.95	0.81
DT	0.77	0.85	0.68	0.80
NN	0.90	0.87	0.92	0.90
SVM	0.90	0.85	0.95	0.90
XGBoost	0.79	0.81	0.75	0.79

TABLE 3. Performance measures obtained from LLM and DT classifiers on the dataset restricted to the first three features. In detail: *back rotation position in sagittal plane, wrist jerk coefficient of variation* and *ChestACCMean* for LLM; *back rotation position in sagittal plane, wrist jerk coefficient of variation* and *HipACCMean* for DT.

Method	Accuracy	Sensitivity	Specificity	F1-score
LLM	0.84	0.73	0.97	0.83
DT	0.76	0.77	0.75	0.79

TABLE 4. Performance measures obtained for LLM with parameters changes.

Configurations		Performance			
Minimized number of conditions	Maximum error allowed (%)	Accuracy	Sensitivity	Specificity	F1-score
False	5	0.84	0.73	0.97	0.83
True	2	0.80	0.69	0.73	0.78
True	10	0.80	0.73	0.88	0.79
False	10	0.84	0.73	0.97	0.83

By looking at the Table 2 for LLM and DT, we can infer that LLM outperforms DT in terms of general classification capabilities, measured by accuracy and F1-score. However, LLM is less sensitive and far more specific than DT. When compared to NN and SVM, LLM and DT show an overall worsened performance (accuracy, F1-score), LLM is high-specific as those black-box models, while DT reaches almost the same sensitivity. XGBoost fits in the middle as regards the accuracy, while has lower F1-score even with respect to LLM and DT.

It is known [70] that a trade-off exists between accuracy and interpretability, so that complex black-box models often outperform interpretable models. Anyway, fatigue detection is a critical field, since it deals with human health: for this reason, we favour explainable models with acceptable performance in spite of the excellent accuracy of black-boxes, which gives no clue on *why* the predictions are built. On that basis, we focused on LLM and DT. In order to assess if a better perfomance could be achieved, we evaluated LLM and DT on the first three most important features obtained from the feature ranking of the previous classification task with default configurations (see figures 6 and 7). In Table 3 we show the obtained metrics in this case. It's possible to see that, with respect to the results in Table 2, the LLM improved in all the indicators, while DT worsened in each metric except for specificity. Again, LLM shows higher specificity whereas DT has higher sensitivity.

Between LLM and DT, the first one globally outperforms the latter. Hence, we also evaluated LLM with different

changes to its default configurations (Table 4). Analyzing the differences between the four tested configurations of LLM parameters, we can observe that such parameters changes actually do not have a significant impact on performance with respect to the default case (Table 2). In particular, when the maximum error allowed is doubled from default (i.e., raised to 10%), performance metrics do not show a corresponding improvement.

For this reason, for further analysis, we decided to adopt the default LLM model, with maximum error allowed of 5% and the minimization of the number of conditions.

B. NON-FATIGUE REGIONS

By considering the default logic learning machine (Table 2), we looked for regions where fatigue was predicted with FNR=0 by implementing the thresholds tuning procedure as explained in sections IV-D2 and IV-D1. For completeness, we remind that FNR is computed as the following ratio: $FNR = \frac{FN}{TP+FN}$.

As regards the approach described in section IV-D1, the first two most important intervals for fatigued class that we got from LLM value ranking were *back rotation position in sagittal plane* > 0.03 and *wrist jerk coefficient of variation* > 0.03. We applied the optimization algorithm (Eq. 8) on such intervals and obtained $\delta^*_{s_1} = -13, \delta^*_{s_2} = 28$. For such values, we got FNR=0 and TNR=0.20. Therefore, the ''non-fatigue region'' 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 resulting 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 (e.g., we can't assume that a subject who stays still won't ever get fatigued). As far as we know, literature does not provide standard ranges and quantifying them is not trivial, requiring specific studies.

To address this issue, we adopted an *a-posteriori* data-driven approach. We evaluated the age histogram (not reported) for the dataset and individuated age ≤ 40 y.o. and age > 40 y.o. as a proper almost balanced split in which to divide test set instances. We computed ranges for *back rotation position in sagittal plane* and *wrist jerk coefficient of variation* by taking their minimum and maximum values within the corresponding age group. 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 follows:

$$\begin{aligned}\mathcal{S}_1 &= f_1 \in (-2.52, 0.42) \\ &\quad \wedge f_2 \in (-1.78, -0.81) \text{ for } age \leq 40\ y.o \\ \mathcal{S}_1 &= f_1 \in (-1.86, 0.42) \\ &\quad \wedge f_2 \in (-2.0, -0.81) \text{ for } age > 40\ y.o\end{aligned}$$

In Figure 2 a visual representation of the obtained regions is provided.

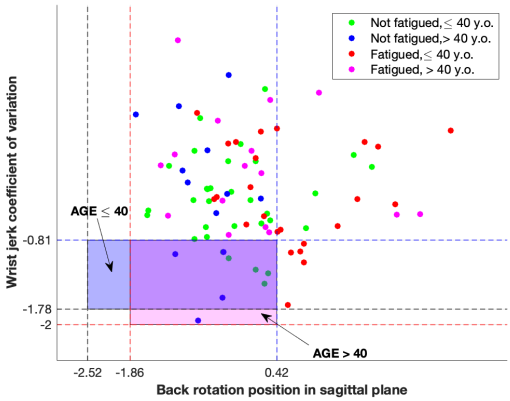

FIGURE 2. Scatter plot of the first two features (back rotation position in sagittal plane and wrist jerk coefficient of variation) with representations of the "non-fatigue region" individuated for age ≤ 40 group (pink) and age > 40 (violet).

In this way, we have exploited fatigued class rules to define a region where the LLM should never predict a non-fatigued state when a subject is actually fatigued.

As described in section IV-D2, we considered the problem of identifying non-fatigue regions starting from the non-fatigued class, too. The value ranking shown *back rotation position in sagittal plane* $<= 0.03$ and *chest acceleration mean* > -0.47 as the two most relevant intervals for predicting non-fatigued class. On such conditions, we applied the optimization problem (eq. 10), which led us to individuate $\delta^*_{t_1} = 79.96$, $\delta^*_{s_2} = 5.71$. 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 previous approach, we limited such region in function of the two group ages (up to and over 40 years old). This procedure redefines $\mathcal{S}_0$ for the age groups as follows (Fig. 3):

$$\begin{aligned}\mathcal{S}_0 &= f_1 \in (-2.52, -1.68) \\ &\quad \vee f2 \in (3.65, 3.99) \textit{ for age} \leq 40\ y.o. \\ \mathcal{S}_0 &= f_1 \in (-1.86, -1.68) \\ &\quad \vee f2 \in (3.65, 3.99) \textit{ for age} > 40\ y.o.\end{aligned}$$

By comparing the results of the application of our two innovative approaches for the individuation of features intervals where the LLM predictions of physical fatigue have FNR=0, we can infer that using the fatigued class (optimization problem 8) allows us to find regions with higher coverage (here intended as the TNR) than using the other approach in eq. 10. This difference is supported by the fact that the second most important feature for LLM is different in the two classes (feature ranking shows *wrist jerk coefficient of variation* for fatigued class, whereas *chest acceleration mean* for non-fatigued). Hence, data distributions for the first two features are different in the two cases too: as depicted

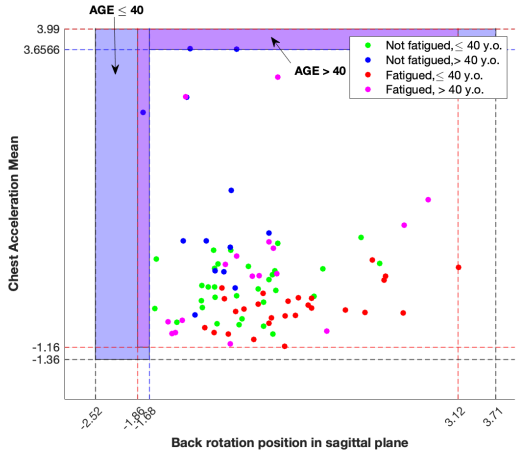

FIGURE 3. Scatter plot of the first two features (back rotation position in sagittal plane, Chest Acceleration Mean) from value ranking of non-fatigued class, with representations of the "non-fatigue regions" based on the age group (violet for age ≤ 40, pink otherwise).

in figures 2 and 3, fatigued and non-fatigued instances are sparser and less separated in the latter, thus explaining the difficulty in individuating wide regions were FNR=0.

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

In order to assess if such value could be increased, we adopted the LLM 0% and built a new predictor as described in section IV-D3. We conducted the optimization process described in equation 12 with $\hat{r}(\boldsymbol{\delta})$ defined by the joining ($\vee$) of a different number, m^0, of rules obtained for the non-fatigued class. $\boldsymbol{\delta}$ is tuned for the first $N^{FR} = 2$ features from non-fatigued feature ranking, namely *HipACCMean* and *WristjerkMean*.

As an example, we report here the two rules for non-fatigued class with highest coverage. The other considered four rules, not reported here for the sake of brevity, present a qualitatively similar structure and deal with the present features, mixed with the others with less score in the ranking.

if (0.51 < HipACCMean ≤ 1.98 **and** ChestACCcoefficientofvariation ≤ 1.11 **and** −1.73 < averagestepdistance ≤ 0.81 **and** backrotationpositioninsagplane ≤ 0.52 **and** −1.42 < WristACCcoefficientofvariation ≤ 1.49 **then non-fatigued** ($C = 52\%$)

if (WristjerkMean > 0.55 **and** −1.35 < Back rotation position in sag plane ≤ 0.04) **then non-fatigued** ($C = 33\%$)

As reported in table 5,before any feature perturbation, we performed a logical OR of a different number of rules (m^0) and computed the corresponding FNR and TNR values. As expected, the TNR values were much higher than with the two previous methods. In order to lower the FNR, we perturbed *HipACCMean* and *WristjerkMean* on the joining of 4 rules (m^0 =4). Thus, we obtained $\delta^*_{s_1}$=1.848 and $\delta^*_{t_2}$=0.027 for such features respectively: these thresholds perturbations brought FNR=0, with TNR=0.42, that is still an appreciable value.

TABLE 5. FNR and TNR values obtained by rule-joining optimization, when the number of joined rules m^0 is varied.

m^0	FNR	TNR
2	0.04	0.64
3	0.04	0.71
4	0.06	0.75
5	0.06	0.84

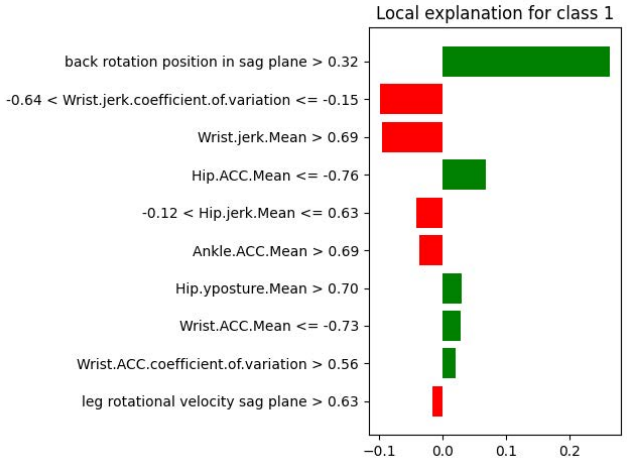

FIGURE 4. Bar plot of LIME explanations for an instance being predicted by the underlying XGBoost as belonging to *fatigued class* with 0.64 prediction probability.

The LLM 0% is therefore able to provide a better performance than the optmized hyper-rectangles above. The conclusion here is consistent with the one in [20] and [23], in which LLM 0% discovered complex safety regions too. It is however worth noting that resorting to simpler schemes, such as the optimization of the hyper-rectangles above, still allows circumvent zero false negatives with acceptable coverage and giving insight into the population of non-fatigued subjects.

C. LIME EXPLANATIONS

After the adoption of SP-LIME with budget $B = 5$, we obtained the set of the 5 most representative LIME explanations of XGBoost predictions.

We report bar plots for two of them in Figures 4 and 5. Green bars indicate positive LIME weights, associated to a positive influence of the corresponding conditions on the prediction probabilities, while red bars indicate negative weights, thus indicating a negative effect on it. Such local explanations can be useful to inspect single results of a classification black box algorithm (XGBoost in our case), thus investigating which factors gave a higher contribution for that particular instance. LIME provides a *local* ranking of the most influent values intervals for the most relevant features: this is the same concept of LLM value ranking (see III-A1), with the difference that for LLM it is *global*.

D. LLM AND DT VERSUS LIME FEATURE RANKING

In Figures 6, 7 and 8, we provide the overall feature ranking bar plots obtained for LLM, DT and LIME respectively. In the latter case, we adopted the approach as described in IV-E.

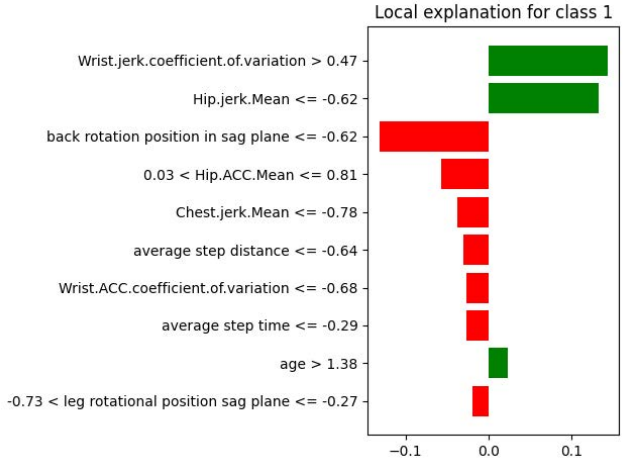

FIGURE 5. Bar plot of LIME explanations for an instance being predicted by underlying XGBoost as belonging to *non-fatigued class* with 0.54 prediction probability.

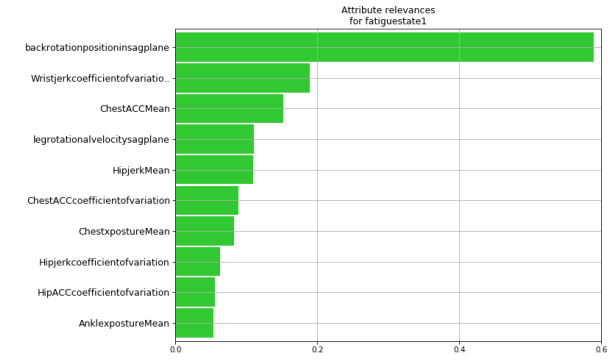

FIGURE 6. Feature ranking for LLM.

All the three rankings individuate the same first two most important features involved in the classification task, that are *back rotation position in sagittal plane* and *wrist jerk coefficient of variation*. This inter-methods concordance suggests that these two features actually have a greater impact in physical fatigue detection than the others.

E. LLM AND DT RULES VERSUS ANCHORS

In the following, as an example, we show the first two rules obtained from LLM and DT without parameters changes for *fatigued* class, ranked by their covering percentage C.

LLM:

if (WristjerkMean $\leq$ 1.22 **and** AnklejerkMean $\leq$ 1.64 **and** AnklexpostureMean $>$ -1.82 **and** legrotationalvelocitysagplane $\leq$ 1.15 **and** backrotationpositioninsagplane $>$ 0.034 **and** WristACCcoefficientofvariation $\leq$ 2.56 **and** Hipyposturecoefficientofvariation $\leq$ 2.77) **then fatigued** ($C = 58\%$)

if (AnklexpostureMean > -1.79 and legrotationalvelocitysagplane $\leq$ 0.46 **and** averageverticalimpact $\leq$ 1.79 **and** Wristjerkcoefficientofvariation $>$ 0.025) **then fatigued** ($C = 45\%$)

DT:

if (Wristjerkcoefficientofvariation $>$ 0.63) **then fatigued** ($C = 36\%$)

if (Wristjerkcoefficientofvariation $\leq$ 0.63 **and** backrotationpositioninsagplane $>$ 0.032 **and** averagesteptime $>$ 0.27) **then fatigued** ($C = 20\%$)

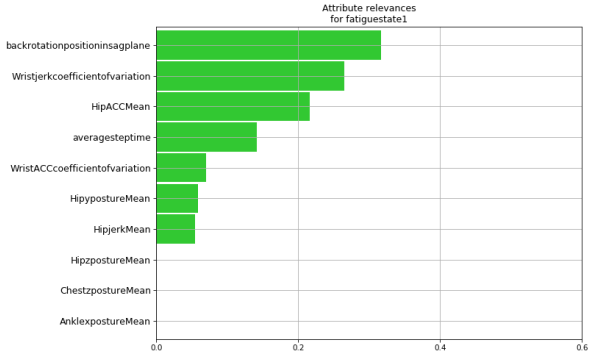

FIGURE 7. Feature ranking for DT.

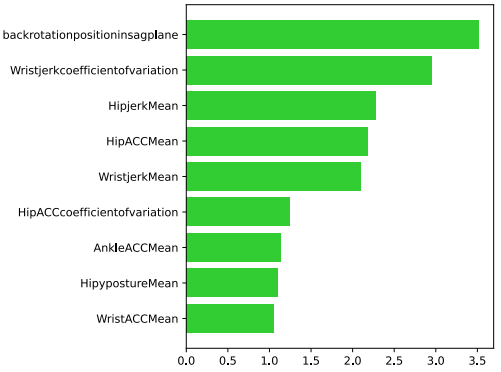

FIGURE 8. Feature ranking for LIME.

TABLE 6. Results of the Anchors optimization problem at varying of the considered Anchor rule. k is an index identifying Anchor a_k; logical OR combinations of single anchors are indicated as $k = 4$, $k = 5$, $k = 6$, referring to $(k = 1 \vee k = 2)$, $(k = 1 \vee k = 3)$, $(k = 1 \vee k = 2 \vee k = 3)$ respectively.

k	$\delta^*_{1_k}$	$\delta^*_{2_k}$	Error	Coverage
1	0.11	1.76	0.04	0.36
2	0.11	0.71	0.04	0.20
3	**1.86**	**0.06**	**0.04**	**0.43**
4	0.11	0.71	0.04	0.38
5	0.66	0.01	0.04	0.41
6	0.11	0.71	0.04	0.43

On the one hand, LLM rules have higher number of conditions reaching higher values of coverage; on the other hand, DT rules contain less conditions but present lower coverage.

The first 3 Anchors obtained for the same class, ranked by Anchors intrinsic definition of coverage C_a, are:

1) **if** (Back rotation position in sag plane $>$ 0.32 **and** Chest.ACC.Mean ≤ -0.25) **then fatigued** ($C_a = 13\%$)
2) **if** (Back rotation position in sag plane $>$ 0.32 **and** Ankle.ACC.coefficient.of.variation $>$ 0.73) **then fatigued** ($C_a = 12, 5\%$)
3) **if** (Back rotation position in sag plane $>$ 0.32 **and** average step time $>$ 0.24) **then fatigued** ($C_a = 11\%$)

When comparing LLM and DT rules with Anchors rules, the most noticeable difference lies in the values of coverage, that are significantly lower for Anchors. This result was what we expected, since Anchors are built upon single instances, whereas LLM and DT consider the whole dataset.

Starting from that, we verified if it was possible to tune the conditions inside such low-coverage rules for fatigued

TABLE 7. Summary of the tests and related results.

Method	Description	Performance
LLM, DT vs NN, SVM, XGBoost	Performance comparison (sensitivity, specificity, accuracy and F1-score) between global transparent-by-design vs black-box models	H igher overall metrics for black box models; LLM outperforms DT (except for sensitivity)
Sensitivity from outside, inside and LLM 0%	Identification of zero FNR "non-fatigue regions" via LLM feature and value rankings	Outside (FNR=0, TNR=0.20) outperforms inside (FNR=0, TNR =0.06); LLM 0% reaches far higher TNR (FNR=0, TNR=0.42)
LLM and DT vs XGBoost+LIME	Comparison of global transparent-by-design models (LLM, DT) and local post-hoc explanations (LIME) of black-box models (XGBoost) through the respective feature ranking (FR). For LLM and DT, FR is directly available; for LIME, an ad-hoc FR was built.	All the three rankings individuated the same first two most important features: back rotation position in sagittal plane and wrist jerk coefficient of variation
LLM and DT vs XGBoost+Anchors	Optimization of Anchors (i.e., rules extracted from single XGBoost predictions) conditions thresholds to extend their covering while maintaining low errors, for comparison with global rule-based models like LLM and DT	In all cases, the covering increased with error fixed to 0.04. In one of the cases, covering raises from 11% to 43%, similar to LLM rules and even higher than covering values for DT

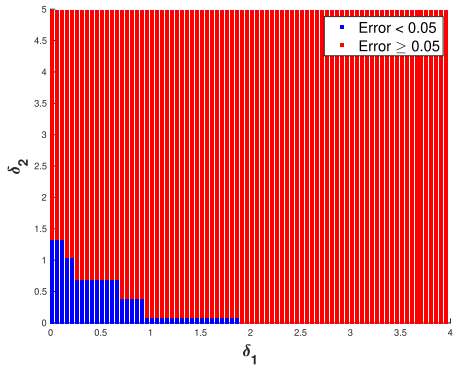

FIGURE 9. Bidimensional plot of $\delta_{1_3}, \delta_{2_3}$ values labelled with respect to the error obtained by applying Anchor $k = 3$ perturbed by their values.

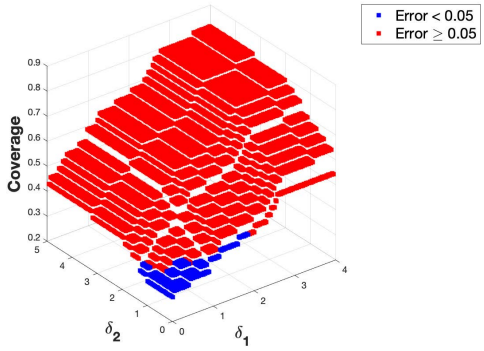

FIGURE 10. Tridimensional plot showing how coverage changes with respect to $\delta_{1_3}, \delta_{2_3}$, inside (blue) and outside (red) the acceptable error threshold.

class by applying the optimization algorithm as explained in Section IV-E. The test was repeated on single Anchors and by joining combinations of them in OR ($\vee$) logical operation to form single rules.

In table 6 are reported the obtained optimal values for $\delta^*_{1_k}$ and $\delta^*_{2_k}$ (the notation is the same adopted in section IV-E). Comparing such values for all the six test cases, we can conclude that the best result is achieved through changes to the third Anchor only (for which $\delta^*_{1_3} = 1.86$ and $\delta^*_{2_3} = 0.06$). With such optimal values, we computed the accuracy (0.70), the specificity (0.95) and F1-score (0.58) too. In this case, coverage reaches 43% as in the case of all the three anchors joined in OR: the choice between these two highest-coverage cases is driven by the fact that in the first one a greater overall expansion of conditions is possible. Furthermore, results show that it's not possible to raise both δ_{1_k} and δ_{2_k} at the same time, while remaining under the acceptable threshold for error.

To further highlight this kind of trade-off between δ_{1_k} and δ_{2_k} for $k = 3$, we plotted different values of δ_{1_3} and δ_{2_3} with blue color if the corresponding error was below 0.05 and red if it was not (Figure 9, the trade-off is visible at the boundary between blue and red zones). Figure 10 shows the coverage function $C(a_3(\delta_{1_3}, \delta_{2_3}))$ again with blue color for acceptable error $E(a_3(\delta_{1_3}, \delta_{2_3}))$ and red for non acceptable. In conclusion, this methodology has allowed us to individuate a new expanded Anchor rule, whose coverage reaches a value (43%) which is almost as high as that of LLM rules and it even overcomes the coverage of DT.

VI. CONCLUSION AND FUTURE WORK

In this work, we studied the physical fatigue detection problem under the explainable artificial intelligence (XAI) paradigm a summary of the approaches and results is available in table 7.

We compared global rule-based models (LLM and DT) with well-established black box models. Geometrically speaking, rule-based models represent the classes through hyper-rectangles in the feature space, resulting in simpler shapes than the ones resulting from more complex black-box models like SVM or NN. For this reason, when classes have very intricate separation profiles, these models may lack in performance if compared to black-box methods: in fact, the latter more complex shape is able to better fit the boundaries between the classes. This is particularly evident for application scenarios, like the presented one, where the available

data are noisy. However, provided to still achieve adequate performance metrics via rule-based models, the advantages of their adoption become evident. Indeed, XAI is useful when one wants to inspect the reasons behind predictions (this is the goal of *explainability*) and how to make such predictions able to guarantee workers safety during occupational tasks and prevent accidents (this is the goal of *responsibility*). Nonetheless, we must remark that achieving such kinds of guarantees, through whatever kind of modelling approach, is still a challenge that requires further work.

By focusing on Responsible AI more in detail, safety has been pursued by introducing innovative rule optimization methodologies (Section IV-D) to let the prediction model guarantee zero statistical error (i.e., the worker is predicted not fatigued without error).

Moreover, we achieved a comparison between global transparent-by-design models (LLM and DT) and local post-hoc model-agnostic explanations models (LIME and Anchors) through novel methodologies. From the comparison of LLM and DT with LIME by means of absolute feature ranking, we found that all the techniques discover the same first two features, thus showing their high impact on our classification task.

Further, our new Anchors optimization method (Section IV-E) has allowed to generalize local rules and make them valid for a larger quantity of data.

The following issues deserve further attention and may constitute future work. Black-box models with subsequent rule extraction is another way to look at the problem under a trade off between accuracy and intelligibility [71], including deep learning [72], [73].

In pursuing the goal of identifying ''non-fatigue regions'', we adopted an *a-posteriori* approach to limit their maximum extension based on age groups after having found the regions. Improvements of our methods may then include an *a-priori* approach instead, starting from two separate datasets related to the age groups: this goal requires larger datasets with balanced age groups. Also, another improvement could rely on the validation of the identified ''non-fatigue regions'' on new data through cross-validation and selective rule extraction [74]. Finally, intersection between machine learning and formal logic to empower the knowledge mining process may be considered too [75].

ACKNOWLEDGMENT

The authors would like to thank the reviewers, for their comments on the proposed AI techniques.

REFERENCES

[1] K. Sadeghniiat-Haghighi and Z. Yazdi, ''Fatigue management in the workplace,'' *Ind. Psychiatry J.*, vol. 24, no. 1, pp. 12–17, 2015. [Online]. Available: https://www.industrialpsychiatry.org/article.asp?

[2] Z. S. 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,'' *Exp. Syst. Appl.*, vol. 155, Oct. 2020, Art. no. 113405. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420302293

[3] T. Kopp, M. Baumgartner, and S. Kinkel, ''Success factors for introducing industrial human–robot interaction in practice: An empirically driven framework,'' *Int. J. Adv. Manuf. Technol.*, vol. 112, nos. 3–4, pp. 685–704, Jan. 2021.

[4] A. Desai, T. Dreossi, and S. A. Seshia, ''Combining model checking and runtime verification for safe robotics,'' in *Proc. Int. Conf. Runtime Verification*, 2017, pp. 172–189.

[5] M. Askarpour, ''Risk assessment in collaborative robotics,'' in *Proc. FMDS*, 2016, pp. 1–6.

[6] *Robots and Robotic Devices—Collaborative Robots*, Standard ISO/TS 15066:2016, 2016. [Online]. Available: https://www.iso.org/standard/62996.html

[7] P. Li, R. Meziane, M. J. Otis, H. Ezzaidi, and P. Cardou, ''A smart safety helmet using IMU and EEG sensors for worker fatigue detection,'' in *Proc. IEEE Int. Symp. Robotic Sens. Environ. (ROSE)*, Oct. 2014, pp. 55–60.

[8] N. Williams, ''The Borg rating of perceived exertion (RPE) scale,'' *Occupational Med.*, vol. 67, no. 5, pp. 404–405, Jul. 2017, doi: 10.1093/occmed/kqx063.

[9] Y. Yu, H. Li, X. Yang, L. Kong, X. Luo, and A. Y. Wong, ''An automatic and non-invasive physical fatigue assessment method for construction workers,'' *Autom. Construction*, vol. 103, pp. 1–12, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518308422

[10] A. Holzinger, ''Explainable AI and multi-modal causability in medicine,'' *i-com*, vol. 19, no. 3, pp. 171–179, Jan. 2021.

[11] C. Rudin, ''Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,'' *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[12] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, ''Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies,'' *Artif. Intell.*, vol. 294, May 2021, Art. no. 103459.

[13] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, ''Explainable AI: A review of machine learning interpretability methods,'' *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[14] A. Adadi and M. Berrada, ''Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),'' *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[15] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, ''Explainable artificial intelligence: Concepts, applications, research challenges and visions,'' in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham, Switzerland: Springer, 2020, pp. 1–16.

[16] D. S. Weld and G. Bansal, ''The challenge of crafting intelligible intelligence,'' *Commun. ACM*, vol. 62, no. 6, pp. 70–79, May 2018.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, ''A survey of methods for explaining black box models,'' *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018, doi: 10.1145/3236009.

[18] 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,'' *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[19] *Road Vehicles Safety of the Intended Functionality*, Standard ISO/PAS 21448:2019, International Organization for Standardization, Geneva, Switzerland, Mar. 2019.

[20] M. Mongelli, E. Ferrari, M. Muselli, and A. Fermi, ''Performance validation of vehicle platooning through intelligible analytics,'' *IET Cyber-Phys. Syst., Theory Appl.*, vol. 4, no. 2, pp. 120–127, Jun. 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cps.2018.5055

[21] M. Mongelli and V. Orani, ''Stability certification of dynamical systems: Lyapunov logic learning machine,'' in *Proc. Int. Conf. Appl. Soft Comput. Commun. Netw. (ACN)*, 2020, pp. 221–235.

[22] M. Mongelli, M. Muselli, A. Scorzoni, and E. Ferrari, ''Accellerating PRISM validation of vehicle platooning through machine learning,'' in *Proc. 4th Int. Conf. Syst. Rel. Saf. (ICSRS)*, Nov. 2019, pp. 452–456.

[23] M. Mongelli, ''Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence,'' *Comput. Commun.*, vol. 179, pp. 166–174, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366421002504

[24] J. Zhang, T. E. Lockhart, and R. Soangra, "Classifying lower extremity muscle fatigue during walking using machine learning and inertial sensors," *Ann. Biomed. Eng.*, vol. 42, no. 3, pp. 600–612, Mar. 2013.

[25] Z. S. Maman, M. A. A. Yazdi, L. A. Cavuoto, and F. M. Megahed, "A data-driven approach to modeling physical fatigue in the workplace using wearable sensors," *Appl. Ergonom.*, vol. 65, pp. 515–529, Nov. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687017300261

[26] A. Baghdadi, F. M. Megahed, E. T. Esfahani, and L. A. Cavuoto, "A machine learning approach to detect changes in gait parameters following a fatiguing occupational task," *Ergonomics*, vol. 61, no. 8, pp. 1116–1129, 2018, doi: 10.1080/00140139.2018.1442936.

[27] J. O. Wobbrock, A. D. Wilson, and Y. Li, "Gestures without libraries, toolkits or training: A $ 1 recognizer for user interface prototypes," in *Proc. 20th Annu. ACM Symp. User Interface Softw. Technol.* New York, NY, USA: Association for Computing Machinery, 2007, pp. 159–168, doi: 10.1145/1294211.1294238.

[28] L. Zhang, M. M. Diraneyya, J. Ryu, C. T. Haas, and E. M. Abdel-Rahman, "Jerk as an indicator of physical exertion and fatigue," *Autom. Construction*, vol. 104, pp. 120–128, Aug. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518310318

[29] S. Lim and C. D'Souza, "Statistical prediction of load carriage mode and magnitude from inertial sensor derived gait kinematics," *Appl. Ergonom.*, vol. 76, pp. 1–11, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687018306690

[30] S. R. Lamooki, J. Kang, L. Cavuoto, F. Megahed, and L. Jones-Farmer, "Challenges and opportunities for statistical monitoring of gait cycle acceleration observed from IMU data for fatigue detection," in *Proc. 8th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechtron. (BioRob)*, Nov. 2020, pp. 593–598.

[31] S. Karvekar, M. Abdollahi, and E. Rashedi, "A data-driven model to identify fatigue level based on the motion data from a smartphone," in *Proc. IEEE Western New York Image Signal Process. Workshop (WNYISPW)*, Oct. 2019, pp. 1–5.

[32] M. Karg, K. Kühnlenz, M. Buss, W. Seiberl, F. Tusker, M. Schmeelk, and A. Schwirtz, "Expression and automatic recognition of exhaustion in natural walking," in *Proc. IADIS Int. Conf. (IHCI)*, Jan. 2008, pp. 165–172.

[33] M. Gholami, C. Napier, A. G. Patiño, T. J. Cuthbert, and C. Menon, "Fatigue monitoring in running using flexible textile wearable sensors," *Sensors*, vol. 20, no. 19, p. 5573, Sep. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/19/5573

[34] L. Nie, X. Ye, S. Yang, and H. Ning, "sEMG-based fatigue detection for mobile phone users," in *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, H. Ning, Ed. Singapore: Springer, 2019, pp. 528–541.

[35] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statist. Surv.*, vol. 16, pp. 1–85, Jan. 2022.

[36] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A Bayesian framework for learning rule sets for interpretable classification," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 2357–2393, 2017.

[37] P. R. Rijnbeek and J. A. Kors, "Finding a short and accurate decision rule in disjunctive normal form by exhaustive search," *Mach. Learn.*, vol. 80, no. 1, pp. 33–62, Jul. 2010.

[38] D. Wei, S. Dash, T. Gao, and O. Gunluk, "Generalized linear rule models," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2019, pp. 6687–6696.

[39] T. Wang, "Multi-value rule sets for interpretable classification with feature-efficient representations," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[40] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1675–1684.

[41] F. Wang and C. Rudin, "Falling rule lists," in *Proc. 18th Int. Conf. Artif. Intell. Statist. (PMLR)*, 2015, pp. 1013–1022.

[42] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning certifiably optimal rule lists for categorical data," 2017, *arXiv:1704.01701*.

[43] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1350–1371, 2015.

[44] T. Wang and Q. Lin, "Hybrid predictive models: When an interpretable model collaborates with a black-box model," *J. Mach. Learn. Res.*, vol. 22, no. 137, pp. 1–38, 2021.

[45] D. Pan, T. Wang, and S. Hara, "Interpretable companions for black-box models," in *Proc. Int. Conf. Artif. Intell. Statist. (PMLR)*, 2020, pp. 2444–2454.

[46] T. Wang and C. Rudin, "Causal rule sets for identifying subgroups with enhanced treatment effects," *Informs J. Comput.*, vol. 34, no. 3, pp. 1626–1643, May 2022.

[47] S. Krishnan and E. Wu, "PALM: Machine learning explanations for iterative debugging," in *Proc. 2nd Workshop Hum. Loop Data Anal.*, May 2017, pp. 1–6, doi: 10.1145/3077257.3077271.

[48] S. K. Biswas, M. Chakraborty, B. Purkayastha, P. Roy, and D. M. Thounaojam, "Rule extraction from training data using neural network," *Int. J. Artif. Intell. Tools*, vol. 26, no. 3, 2017, Art. no. 1750006.

[49] C. He, M. Ma, and P. Wang, "Extract interpretability-accuracy balanced rules from artificial neural networks: A review," *Neurocomputing*, vol. 387, pp. 346–358, Apr. 2020.

[50] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy, "TreeView: Peeking into deep neural networks via feature-space partitioning," 2016, *arXiv:1611.07429*.

[51] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1803–1831, 2010.

[52] P. Cortez and M. J. Embrechts, "Opening black box data mining models using sensitivity analysis," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2011, pp. 341–348.

[53] Y. Liang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," *Neurocomputing*, vol. 419, pp. 168–182, Jan. 2021.

[54] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies," *Artif. Intell.*, vol. 294, May 2021, Art. no. 103459. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000102

[55] Y. Bai, Y. Guan, and W.-F. Ng, "Fatigue assessment using ECG and actigraphy sensors," in *Proc. Int. Symp. Wearable Comput.*, Sep. 2020, pp. 12–16.

[56] M. J. Pinto-Bernal, C. A. Cifuentes, O. Perdomo, M. Rincón-Roncancio, and M. Múnera, "A data-driven approach to physical fatigue management using wearable sensors to classify four diagnostic fatigue states," *Sensors*, vol. 21, no. 19, p. 6401, Sep. 2021.

[57] M. R. Al-Mulla and F. Sepulveda, "Novel feature modelling the prediction and detection of sEMG muscle fatigue towards an automated wearable system," *Sensors*, vol. 10, no. 5, pp. 4838–4854, May 2010, doi: 10.3390/s100504838.

[58] A. Kozarev, J. Quindlen, J. How, and U. Topcu, "Case studies in data-driven verification of dynamical systems," in *Proc. 19th Int. Conf. Hybrid Syst., Comput. Control (HSCC)*, Apr. 2016, pp. 81–86.

[59] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[61] M. Muselli, "Switching neural networks: A new connectionist model for classification," in *Proc. Int. Workshop Natural Artif. Immune Syst.*, Jan. 2005, pp. 23–30.

[62] S. Parodi, C. Dosi, A. Zambon, E. Ferrari, and M. Muselli, "Identifying environmental and social factors predisposing to pathological gambling combining standard logistic regression and logic learning machine," *J. Gambling Stud.*, vol. 33, no. 4, pp. 1121–1137, Dec. 2017.

[63] S. Parodi, C. Manneschi, D. Verda, E. Ferrari, and M. Muselli, "Logic learning machine and standard supervised methods for Hodgkin's lymphoma prognosis using gene expression data and clinical variables," *Health Informat. J.*, vol. 24, no. 1, pp. 54–65, Mar. 2016.

[64] S. Parodi, R. Filiberti, P. Marroni, R. Libener, G. P. Ivaldi, M. Mussap, E. Ferrari, C. Manneschi, E. Montani, and M. Muselli, "Differential diagnosis of pleural mesothelioma using logic learning machine," *BMC Bioinf.*, vol. 16, no. S9, pp. 1–10, Dec. 2015.

[65] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[66] M. Mongelli, M. Muselli, and E. Ferrari, "Achieving zero collision probability in vehicle platooning under cyber attacks via machine learning," in *Proc. 4th Int. Conf. Syst. Rel. Saf. (ICSRS)*, Nov. 2019, pp. 41–45.
[67] C. Molnar. (2019). *Interpretable Machine Learning*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[68] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[69] J. Nobre and R. F. Neves, "Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets," *Exp. Syst. Appl.*, vol. 125, pp. 181–194, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419300995
[70] A. Saranti, B. Taraghi, M. Ebner, and A. Holzinger, "Property-based testing for parameter learning of probabilistic graphical models," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction* in Lecture Notes in Computer Science, vol. 12279, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. R. Weippl, Eds. Dublin, Ireland: Springer, Aug. 2020, pp. 499–515, doi: 10.1007/978-3-030-57321-8_28.
[71] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, *Rule Extraction From Support Vector Machines: An Overview of Issues and Application in Credit Scoring*. Berlin, Germany: Springer, 2008, pp. 33–63 doi: 10.1007/978-3-540-75390-2_2.
[72] T. Hailesilassie, "Rule extraction algorithm for deep neural networks: A review," 2016, *arXiv:1610.05267*.
[73] M. Chakraborty, S. K. Biswas, and B. Purkayastha, "Rule extraction from neural network trained using deep belief network and back propagation," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3753–3781, Sep. 2020.
[74] D. Cangelosi, F. Blengio, R. Versteeg, A. Eggert, A. Garaventa, C. Gambini, M. Conte, A. Eva, M. Muselli, and L. Varesio, "Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients," *BMC Bioinf.*, vol. 14, no. 7, Apr. 2013, pp. 1–20. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1040674960 and https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-14-S7-S12
[75] R. Evans and E. Grefenstette, "Learning explanatory rules from noisy data," *J. Artif. Intell. Res.*, vol. 61, pp. 1–64, Jan. 2018.

SARA NARTENI received the M.Sc. degree in bioengineering from the University of Genoa, in March 2020. She is currently pursuing the Ph.D. degree with the Politecnico di Torino. She is also working at the IEIIT Institute of Consiglio Nazionale delle Ricerche. Her M.Sc. thesis titled "Pleural line ultrasound videos analysis for computer aided diagnosis in acute pulmonary failure." She works on data analytics and machine learning topics from different fields, such as industry, healthcare, and automotive. Her research interests include computer security topics, including covert channels and the Internet of Things.

VANESSA ORANI received the M.Sc. degree in stochastics and data science, in April 2019. Her M.Sc. thesis "Bayesian isotonic logistic regression via constrained splines: an application to estimate the serve advantage in professional tennis." She is currently a Researcher/a Developer at the Laboratory of Aitek S.p.A. Her research interests include machine learning applied in different field, including transport, telecommunications, and video surveillance. She is currently involved in a project to investigate AI approaches in video content analysis/image processing.

ENRICO CAMBIASO received the Ph.D. degree in computer science from the University of Genoa. He is currently working at Ansaldo STS and Selex ES, both companies are part of the Finmeccanica Group. He has a strong background as a Computer Scientist and he is also employed at the IEIIT Institute of Consiglio Nazionale delle Ricerche, as a Technologist working on cyber-security topics and focusing on the design of last generation threats and related protection.

MATTEO RUCCO received the master's degree in data science for complex system modeling and analysis and the Ph.D. degree in information science and complex systems. During the Ph.D. degree, he defined a new data-driven methodology that combines topology, information theory (i.e., persistent entropy), and automata theory for modeling complex systems behavior. He is currently a Data Scientist with ten years of experience in data science for complex bio-inspired systems, aerospace, elevators, cyber-physical systems, and medical industries. His publication list counts more than ten patents applications and more than 30 research papers. His research interest includes defining new data-driven methodologies for making artificial intelligence-based systems trustworthy. Currently, he is the Coordinator of the 1-SWARM Horizon 2020 EU Project.

MAURIZIO MONGELLI (Member, IEEE) received the Ph.D. degree in electronics and computer engineering from the University of Genoa (UNIGE), in 2004. His Ph.D. degree was funded by Selex Communications S.p.A. (Selex). He worked for both Selex and the Italian Telecommunications Consortium (CNIT), from 2001 to 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, Munich. He is currently a Researcher at the Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), National Research Council (CNR), where he deals with machine learning applied to bioinformatics and cyber-physical systems. He is the coauthor of over 100 international scientific papers, two patents and is participating in the SAE G-34/EUROCAE WG-114 AI in Aviation Committee.