A New XAI-based Evaluation of Generative Adversarial Networks for IMU Data Augmentation

A New XAI-based Evaluation of Generative Adversarial Networks for IMU Data Augmentation

Sara Narteni
Consiglio Nazionale delle Ricerche (CNR)
IEIIT institute
Genoa, Italy
Politecnico di Torino
DAUIN Department
sara.narteni@ieiit.cnr.it

Vanessa Orani
Aitek S.p.A.
Genoa, Italy
vorani@aitek.it

Enrico Ferrari
Rulex Innovation Labs
Genoa, Italy
enrico.ferrari@rulex.ai

Damiano Verda
Rulex Innovation Labs
Genoa, Italy
damiano.verda@rulex.ai

Enrico Cambiaso
Consiglio Nazionale delle Ricerche (CNR)
IEIIT institute
Genoa, Italy
enrico.cambiaso@ieiit.cnr.it

Maurizio Mongelli
Consiglio Nazionale delle Ricerche (CNR)
IEIIT institute
Genoa, Italy
maurizio.mongelli@ieiit.cnr.it

***Abstract*—Data augmentation is a widespread innovative technique in Artificial Intelligence: it aims at creating new synthetic data given an existing real baseline, thus allowing to overcome the issues arising from the lack of labelled data for proper training of classification algorithms. Our paper focuses on how a common data augmentation methodology, the Generative Adversarial Networks (GANs), which is widespread for images and time-series data, can be also applied to generate multivariate data. We propose a novel scheme for GANs evaluation, based on the performance of an explainable AI (XAI) algorithm and an innovative definition of rule similarity. In particular, we will consider an application dealing with the augmentation of Inertial Movement Units (IMU) data for physical fatigue monitoring in two age subgroups (under and over 40 years old) of the original data. We will show how our innovative rule similarity metric can drive the selection of the best fake dataset among a set of different candidates, corresponding to different GAN training runs.**

***Index Terms*—Data augmentation, Generative Adversarial Networks, Explainable AI, Rule similarity**

I. Introduction

Artificial Intelligence (AI) methodologies are spreading in many socioeconomic fields today, including healthcare. In this context, AI represents a promising tool for clinical decision support systems (CDSS) in diagnostic/prognostic processes, health status monitoring [1], drug discovery, public health management and many other applications [2]. There are different actors involved in this effort towards the automation of healthcare, ranging from AI experts who develop the methodologies to the clinicians/health systems that adopt the solutions and, finally, to the patients. Smart solutions are offered thanks to the Internet of Things and the advancements in deep learning [3], but there is still need to open the "black box", thus allowing each actor to understand the logic behind the algorithms. Moreover, errors on algorithmic predictions must be avoided in healthcare, since it is a safety-critical context. As a consequence, AI research is recently addressing its interests on explainable AI (XAI), which consists in providing AI solutions that are understandable even by non-expert users, also in compliance with recent legislation [4], like the European GDPR (https://gdpr.eu/tag/gdpr/). XAI is also an important tool for error control. There are many different methods falling under XAI definition today [5].However, developing accurate XAI-based (and, more in general, AI-based) CDSS is often affected by the quality of the available datasets, which are usually incomplete or too complex and huge [6].

Data augmentation, whose aim is to create synthetic datasets based on the available real data, represents a solution to this issue.

In this work, we extend a common data augmentation method for images and time series, i.e. the Generative Adversarial Networks, to multivariate data and present an innovative evaluation method for the obtained synthetic datasets, based on the performance of a rule-based XAI algorithm (the Logic Learning Machine) for a classification task and a new measure of rule similarity.

II. Related Work

Data augmentation is an emerging process aiming at the creation of fake data based on samples of real data. There is a variety of techniques that have been developed for different kinds of healthcare data. Surely, the most common field where data augmentation is successfully applied is biomedical imaging: in this context, new fake images can be created by modifying the geometry of the original ones or by adding noise [7], [8]. However, this group of techniques tend to generate synthetic images that have similar distributions to the real ones: this is a limitation, since data augmentation

is often used to improve machine/deep learning performance. For this reason, another widespread technique is the use of Generative Adversarial Networks (GANs) for different clinical purposes [9]–[11]. GANs are also explored for time-series augmentation. In [12] they are applied to create artificial raw audio of cough to help respiratory disease diagnosis. In [13], GANs for ECG and EEG signals classification are built using Recurrent Neural Networks with LSTM hidden layers for both generator and discriminator. Data augmentation techniques based on GANs for multivariate data (among others) are included in [14]. Moreover, in [15], corGAN aims to create synthetic health record through a combination of convolutional GANs e convolutional autoencoders; such system was tested on MIMIC-III and UCI seizure epileptic recognition datasets and is evaluated by comparing ML classification performance. MedGAN is proposed in [16], being a combination of GANs and autoencoders to create synthetic Electronic Health Records items, in a privacy-preserving fashion: the evaluation is based on distribution statistics, predictive modeling tasks and review by experts.

In [17], a first evaluation of GAN results based on XAI and rules validation through Fisher test is proposed for the augmentation of small pulmonary syndromes monitoring datasets. To the best of our knowledge, our paper contributes to this field by combining in a systematic way GANs and XAI tools (performance evaluation and rule similarity) for the augmentation of wearable sensors data to detect physical fatigue.

III. Generative Adversarial Networks

Generative Adversarial Networks (GANs) have been introduced with the purpose of generating new, synthetic data (data augmentation) [18]. Starting from the distribution of the training set and a noise distribution, they are able to generate new never seen before data. They produce the whole output all at once, differently from other types of generative models, such as recurrent NNs that generate one element at a time.

The algorithmic principle at the basis of GANs training relies on game theory. The game is posed between a generator network and a discriminator network. The generator uses the encoder–decoder scheme [19] to build artificial data. The discriminator learns the boundary between real and synthetic data. More specifically, the discriminator is trained to become better at distinguishing real from synthetic data; the generator learns to synthesize better data to fool the discriminator (for this reason, the "adversarial" term).

In particular, we implemented a conditional GAN to take into account the class labels, where both the generator and discriminator are forced to generate new data subject to a condition [20]. The condition is in the form of a one-hot-encoded vector version of the classes:

- In the generator, we associate the noise to a particular class;
- In the discriminator, we classify the samples as real or fake based on real and fake data and their corresponding labels.

Training the discriminator and generator in conditional GANs is similar to training a discriminator and generator in a simple GAN. The only difference is that both the generated fake and real are conditioned with their corresponding one-hot labels. Figure 1 shows the architecture of the implemented conditional GAN [17].

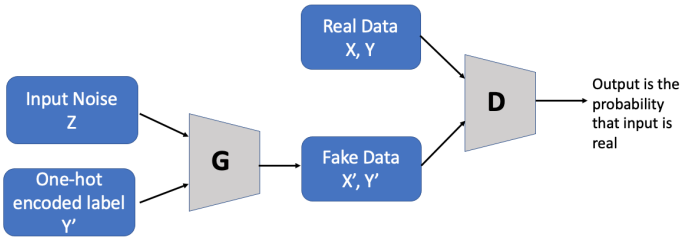

Fig. 1: Conditional GAN scheme.

The optimization scheme of the training is formulated in order to achieve, at convergence, a game equilibrium in which the generator's samples are indistinguishable from real data.

IV. Logic Learning Machine

Logic Learning Machine (LLM) is a supervised model developed by Rulex [21], as the efficient implementation of Switching Neural Networks [22].

Considering classification tasks, the LLM learns a ruleset, made up of a number M of **if-then** rules $\mathbf{r}_k, k = 1, \ldots, M$, each predicting an output class label based on the logical product of their conditions c_{l_k}, $l_k = 1_k, \ldots, d_k$ on input features. Two performance metrics can be derived for each rule, the covering $C(\mathbf{r}_k)$ and the error $E(\mathbf{r}_k)$, defined as follows:

$$C(\mathbf{r}_k) = \frac{TP(\mathbf{r}_k)}{TP(\mathbf{r}_k) + FN(\mathbf{r}_k)} \quad , \quad E(\mathbf{r}_k) = \frac{FP(\mathbf{r}_k)}{TN(\mathbf{r}_k) + FP(\mathbf{r}_k)}, \quad (1)$$

where $TP(\cdot), FP(\cdot), TN(\cdot), FN(\cdot)$ are the confusion matrix values obtained by classifying the samples with the rule. Both covering and error are the basis to define *feature ranking* and *value ranking*. *Feature ranking* orders the variables based on a measure of their relevance in predicting the output class, which depends on the covering and on the increase in the error when a condition is removed from the rule (see Eq. 4 in [23]). *Value ranking* instead individuates the intervals of values, for each attribute, that impact more on the rules output class.

V. Rule Similarity

Whenever it is applied on a dataset, a rule-based model provides a set of rules. Rule similarity consists in defining a measure able to express how much two rules are similar. This can be done with either a semantic or a statistical approach: in the first case, rule similarity is defined based on the information contained into the rules, just like a human can understand it; the latter approach is based on data instances covered by the rule instead.

For sake of simplicity and intepretability, in this paper we develop and adopt the semantic approach to build a measure of

rule similarity for the LLM. Given two sets M_1 and M_2 of m_1 and m_2 rules produced by the LLM (the adopted notation is the same as in Section IV), let us consider two rules $r_k \in M_1$ and $r_z \in M_2$. A condition in r_k is $c_{l_k},\ l_k = 1_k, \ldots, d_k$, and it is associated to a weight w_{l_k}, which is related to the increase in the error of the rule due to the removal of its condition c_{l_k}. Moreover, such condition defines a domain D_{l_k} in the feature space, corresponding to an interval for ordinal variables (or sets of values for nominal variables). Let us now denote with $|D_{l_k}|$ the size of such domain, which is intended as the Euclidean distance for ordinal variables (or cardinality for categorical). Similarly, a condition $c_{i_z},\ i_z = 1_z, \ldots, n_z$ in rule r_z has a weight, w_{i_z}, and defines a domain D_{i_z} of size $|D_{i_z}|$.

Being X_j the attribute associated to condition c_{l_k} of rule , we now define the following binary quantity to express if the attributes of conditions $c_{l_k},\ c_{i_z}$ are the same or not:

$$\beta(X_j(c_{l_k}, c_{i_z})) = \begin{cases} 1 & \text{if } X_j(c_{l_k}) = X_j(c_{i_z}) \\ 0 & \text{if } X_j(c_{l_k}) \neq X_j(c_{i_z}) \end{cases} \quad (2)$$

The similarity between the conditions $c_{l_k},\ c_{i_z}$ takes into account the overlapping of their domains and is then defined as:

$$s(c_{l_k}, c_{i_z}) = \beta(X_j(c_{l_k}, c_{i_z})) \cdot \frac{|D_{l_k} \cap D_{i_z}|}{\max(|D_{l_k}|, |D_{i_z}|)} \quad (3)$$

being $s(c_{l_k}, c_{i_z})$ =0 if the attributes are not the same or the domains do not overlap; if the conditions are identical, it will be $s(c_{l_k}, c_{i_z})$ =1.

Based on this, the similarity between two rules is computed as follows:

$$S(r_k, r_z) = \frac{\sum_{l_k=1}^{d_k} \sum_{i_z=1}^{n_z} s(c_{l_k}, c_{i_z})(w_{l_k} + w_{i_z})}{\sum_{l_k=1}^{d_k} w_{l_k} + \sum_{i_z=1}^{n_z} w_{i_z}} \quad (4)$$

Such equation provides the value $S(r_k, r_z) \in [0, 1]$, being $S(r_k, r_z) = 0$ if the rules do not contain any couple of conditions so that $s(c_{l_k}, c_{i_z})$ >0 and a non-zero weight, $S(r_k, r_z)$ <1 if there is at least a different (or an additional) condition and $S(r_k, r_z)$ =1 if all the conditions in the two rules are identical. Moreover, the value of $S(r_k, r_z)$ can be computed with respect to the same output value of the rules.

Just like the LLM, rule similarity is implemented in Rulex platform too. For practical applications (as in the scope of this work), it may be needed to compare two rulesets in terms of rule similarity, by aggregating the values for different couples of rules, as explained in the following Section V-A.

A. Similarity between rulesets

Considering two different rulesets M_1 and M_2, we want to achieve a single value of rule similarity to compare them. Let us suppose M_1^0 and M_2^0 being the subsets of the rulesets M_1 and M_2 respectively, referring to the output value $y = 0$ of the rules. Similarly, we denote with M_1^1 and M_2^1 the subsets for output class $y = 1$.

First of all, we compute the rule similarities as in Equation V, thus obtaining all the comparisons between M_1^0 and M_2^0, on the one hand, and between M_1^1 and M_2^1 on the other hand. This procedure leads to two sets R^0 and R^1 of N^0 and N^1 non-zero rule similarities for output values $y = 0$ and $y = 1$ respectively. Then, we first compute the arithmetic means $\bar{S^0}$ and $\bar{S^1}$ of the elements in R^0 and R^1, which provide a single value for rule similarity in the corresponding output class. Eventually, the similarity between rulesets M_1 and M_2 is computed again with the arithmetic mean as follows:

$$\bar{S}(M_1, M_2) = \frac{\bar{S^0} + \bar{S^1}}{2} \quad (5)$$

Again, $\bar{S}(M_1, M_2) \in [0, 1]$.

VI. Proposed Evaluation Framework

The overall idea of our proposed XAI-based evaluation of GANs is depicted in Fig. 2. Our purpose is to use explainable models in order to evaluate and understand the behavior and the potential of GANs for classification tasks. In particuar, we adopt the LLM as the XAI model.

First, we perform a baseline assessment by training that model on the real dataset, thus deriving the real ruleset. Then, we perform N runs of data augmentation through GANs, by leaving the parameters (number of epochs, number of hidden layers, number of neurons per hidden layer, learning rate and batch size) fixed at each iteration and by varying only the random input noise to the generator. Hence, N fake datasets are obtained. The next step involves the training of the XAI model both on each generated fake dataset only and on the combination of each fake dataset with the real data: hence, we deal with N "fake rulesets" for the first case and N "real+fake rulesets" for the latter. We then propose three testing scenarios for these rulesets:

- Scenario 1: each "fake ruleset" tested on real data only;
- Scenario 2: each "real+fake ruleset" tested on the concatenation of real and fake data;
- Scenario 3: each "fake ruleset" tested on the concatenation of real and fake data.

In literature, GANs are mainly used in Scenario 2, since data augmentation simply represents a technique to improve models performance. In our case, we introduce scenarios 1 and 3 to understand how GANs work and how they approximate the distribution of the real data. Each scenario is evaluated by measuring the corresponding ruleset's classification performance through two common metrics, i.e. accuracy and F1-score. Using this approach, we are not interested in the convergence of the training algorithm, but we focus on finding the most appropriate augmented dataset generated by different GAN runs via LLM. Moreover, for each GAN trial, we also computed the Fréchet Inception Distance (FID) score [24] as it is the standard measure of GANs results: lower FID values mean better quality and diversification of the generated data.

Based on the metrics values resulting from the different scenarios, we can select the best-performing one and compare the corresponding rulesets with the real ruleset by computing

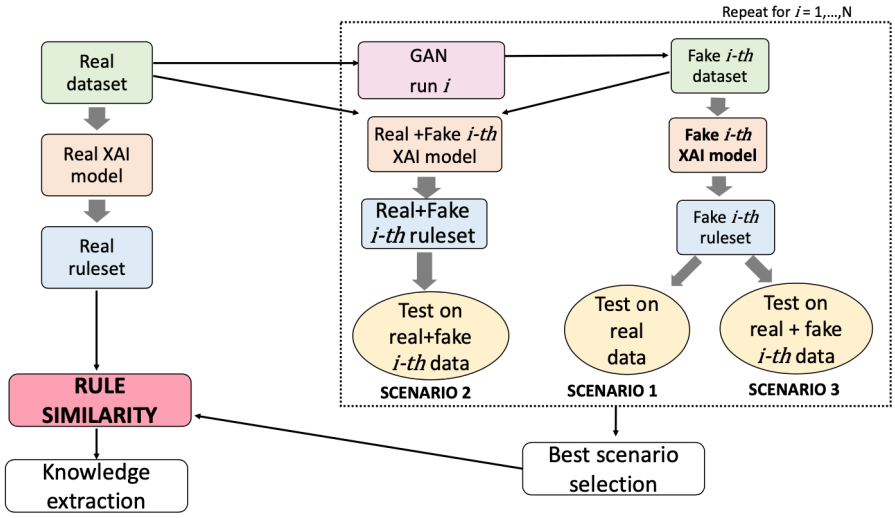

Fig. 2: The proposed GAN evaluation scheme

the rule similarity as described in Sec. V. The results of rule similarity can provide useful information about how data augmentation influences the learning of the rules with respect to the real baseline, thus allowing new knowledge extraction.

VII. Application and Results

A. Physical fatigue dataset

Our application use case deals with an open-source dataset [25] on fatigue detection during the execution of a physical task (Manual Material Handling [26]). Data were collected through wearable Inertial Movement Units (IMUs) from 15 participants who were asked to perform the required task for 180 minutes and provide a fatigue level every 10 minutes using the well-known Borg scale [27]. Fatigued state (class $y = 1$) corresponds to RPE$\geq$13, whereas lower values denote non-fatigued (class $y = 0$). From sensors raw data, a list of features is derived (see Table 2 in [26] for details).

The original 269 samples of the dataset were split into 161 samples corresponding to subjects with $age \leq 40$ years old (under 40 in the rest of the paper) and 108 samples for the $age > 40$ years old group (over 40 in the rest of the paper). We first applied the LLM model to such groups with a 70%/30% for training and test set respectively: the classification results on test set showed an accuracy of 0.63 and F1-score of 0.57 for under 40; for the over 40, the accuracy was 0.84 and F1-score was 0.80. Hence, the under 40 LLM performance was poor with respect to the other age group.

Therefore, we exploited the data augmentation through GANs for creating wider (400 samples) and balanced in size datasets for both the age groups.

B. GAN runs evaluation

For both under 40 and over 40 groups, the GAN training was repeated for $N = 10$ times with the following parameters:

- Generator hidden layers sizes: 128, 64, 32, 1;
- Discriminator hidden layers sizes: 32, 64, 128;
- Batch size: 64;
- Epochs: 5000;
- Learning rate: 5e-5;

The obtained results for each run, in terms of accuracy and F1-score (also denoted with F1 in the Tables), in the three scenarios (as defined in section VI) are shown in Table I for under 40 and Table II for over 40. Besides the metrics related to the rulesets performance, we also show the values of FID for each run (FID involves a comparison between real and fake data distribution, hence it does not depend on the scenarios).

	Scenario 1		Scenario 2		Scenario 3		
Run	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**FID**
1	0.47	0.30	0.81	0.78	0.78	0.75	3654
2	0.49	0.45	0.77	0.76	0.76	0.79	9781
3	0.60	0.63	0.79	0.81	0.81	0.82	17019
4	0.62	0.61	0.66	0.60	0.77	0.77	4237
5	0.60	0.61	**0.86**	**0.86**	0.84	0.84	39580
6	0.58	0.61	0.83	0.83	0.83	0.83	118180
7	0.47	0.23	0.78	0.79	0.79	0.79	38513
8	0.53	0.48	0.75	0.74	0.80	0.79	41702
9	0.50	0	0.59	0.64	0.73	0.79	35786
10	0.58	0.54	0.80	0.78	0.80	0.81	8074

TABLE I: Results of GAN evaluation through LLM in under 40 group

	Scenario 1		Scenario 2		Scenario 3		
Run	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**FID**
1	0.45	0.15	0.68	0.68	0.77	0.74	2393
2	0.53	0.14	0.83	0.82	0.84	0.82	6918
3	0.64	0.69	**0.90**	**0.89**	0.88	0.89	8067
4	0.68	0.71	0.85	0.86	0.89	0.90	41279
5	0.58	0.67	0.84	0.85	0.86	0.87	6832
6	0.54	0.44	0.82	0.80	0.82	0.80	19413
7	0.55	0.65	0.82	0.81	0.83	0.84	10299
8	0.52	0.10	0.68	0.62	0.78	0.75	60261
9	0.64	0.58	0.85	0.85	0.88	0.88	833
10	0.66	0.65	0.85	0.87	0.88	0.89	3939

TABLE II: Results of GAN evaluation through LLM in over 40 group

Observing the obtained values of FID, for both under and over 40, we can infer that in this case we do not achieve consistency between it and accuracy/F1-score metrics: the lowest FID value does not correspond to their highest values. However, there is concordance between the worst accuracy, F1-score and highest FID for run 7 in all the three scenarios.

Focusing on accuracy and F1-score, the immediate consideration following our results is that the evaluation in Scenario 1 does not work neither for under 40 nor for over 40 (low values of accuracy and F1-score in all the runs): therefore, we can not adopt the "fake LLM" rules to improve the performance of the original rules on real datasets only. However, in general, the performance metrics are sensitively higher for Scenario 2 and 3. Considering Scenario 2, the good results we obtained suggest that we can merge real data with synthetic data to improve the original performance of LLM, so we select it as the best Scenario.

C. Rule similarity results

In order to assess which of the 10 fake datasets is the best one, we applied the rule similarity (Section V) approach. In particular, we computed the rule similarity between the real ruleset and the rulesets obtained on real+fake data. Starting from all the obtained non-zero similarities, we aggregated them as explained in Section V-A. The final rule similarity values $\bar{S}$ for each couple of real and real+fake dataset are reported in Table III.

Run	Under 40	Over 40
1	0.17	0.19
2	**0.29**	**0.43**
3	0.24	**0.17**
4	0.20	0.41
5	0.21	0.26
6	0.22	0.40
7	**0.12**	0.30
8	0.16	0.24
9	0.22	0.27
10	0.21	0.30

TABLE III: Rule similarities $\bar{S}$ for under 40 and over 40 groups. Each run individuates a fake dataset, which is combined with the real one. The rule similarities are then computed between the real dataset and each real+fake dataset.

Since we expect that the data augmentation approach might be able to both reproduce quite similar data to the original and also highlight different characteristics that may be important for the analysis, we are interested in the two fake datasets that, when concatenated with the real data, present, respectively, the maximum and the minimum rule similarity with respect to the real ruleset. Referring to Table III, they correspond to run 2 and run 7 for under 40, while run 2 and run 3 for over 40.

Since the obtained rule similarities are all lower than 0.50 for both age groups, and considering the good performance metrics achieved (Tables I and II), we can infer that our application of data augmentation to physical fatigue results in new datasets containing quite different but performing information.

In order to inspect the reasons behind the rule similarity values, we report, as an example, the rulesets corresponding to real over 40 data and real+fake data from run 2 for over 40, for which the rule similarity is the highest ($\bar{S}$ =0.43). The rules obtained with real data were the following:

1) **if** Hip.ACC.Mean $>$ 3.961156 *and* Hip.yposture.Mean $\leq$ 65.652682 *and* 'back rotation position in sag plane' $\leq$ 11.700000 *and* Wrist.jerk.coefficient.of.variation $\leq$ 135.510000 *and* Wrist.ACC.coefficient.of.variation $\leq$ 41.240000 **then non-fatigued** ($C_a = 0.944444$)
2) **if** Ankle.xposture.Mean $\leq$ -5.042815 **then non-fatigued** ($C_a = 0.138889$)
3) **if** Hip.ACC.Mean $\leq$ 3.961156 **then fatigued** ($C_a = 0.575000$)
4) **if** Hip.yposture.Mean $>$ 63.694615 *and* -5.555916 $<$ Ankle.xposture.Mean $\leq$ 4.061175 *and* Chest.jerk.coefficient.of.variation $>$ 79.195000 **then fatigued** ($C_a = 0.525000$)
5) **if** Hip.ACC.Mean $\leq$ 4.609755 *and* Wrist.jerk.coefficient.of.variation $>$ 107.135000 **then fatigued** ($C_a = 0.425000$)

In contrast, when we applied the LLM to the real data joined with the data generated by GAN at run 2, we obtained the following rules:

1) **if** Wrist.jerk.Mean $>$ 13.284318 *and* Ankle.jerk.Mean $>$ 40.381465 *and* 'back rotation position in sag plane' $\leq$ 9.685000 *and* Hip.ACC.coefficient.of.variation $\leq$ 53.000000 **then non-fatigued** ($C_a = 0.695187$)
2) **if** Hip.ACC.Mean $>$ 4.083357 *and* Hip.yposture.Mean $\leq$ 67.313572 *and* Chest.zposture.Mean $>$ 5.373362 *and* 1.309198 $<$ 'time bent' $\leq$ 10.230000 *and* 'average back bent angle' $\leq$ 30.528813 *and* 'mean hip osicllation' $\leq$ 0.216931 *and* 'average vertical impact' $\leq$ 24.647082 *and* Wrist.jerk.coefficient.of.variation $\leq$ 135.510000 *and* Chest.yposture.coefficient.of.variation $>$ 20.742683 **then non-fatigued** ($C_a = 0.684492$)
3) **if** 'average vertical impact' $\leq$ 20.718409 *and* Chest.jerk.coefficient.of.variation $>$ 86.941708 *and* Ankle.jerk.coefficient.of.variation $\leq$ 111.975000 *and* Hip.yposture.coefficient.of.variation $>$ 7.029767 *and* 18.205000 $<$ Chest.yposture.coefficient.of.variation $\leq$ 87.535000 **then non-fatigued** ($C_a = 0.299465$)
4) **if** 'number of steps' $>$ 68.508728 *and* 'average step time' $\leq$ 0.882425 **then non-fatigued** ($C_a = 0.101604$)
5) **if** Wrist.jerk.Mean $\leq$ 13.342289 *and* Hip.ACC.Mean $\leq$ 4.767478 *and* Chest.jerk.Mean $\leq$ 17.961010 *and* Ankle.jerk.Mean $\leq$ 67.325729 *and* 'mean hip osicllation' $>$ 0.060994 *and* Hip.yposture.coefficient.of.variation $\leq$ 26.770000 **then fatigued** ($C_a = 0.704142$)
6) **if** Hip.yposture.Mean $>$ 65.506763 *and* Ankle.xposture.Mean $>$ -6.340669 *and* 'mean hip osicllation' $\leq$ 0.198853 *and* 'back rotation position in sag plane' $>$ 5.319703 **then fatigued** ($C_a = 0.497041$)
7) **if** Wrist.jerk.Mean $\leq$ 14.225740 *and* Chest.yposture.Mean $\leq$ 63.875630 *and* 'average back bent angle' $>$ 19.558336 **then fatigued** ($C_a = 0.313609$)
8) **if** -32.848320 $<$ Hip.xposture.Mean $\leq$ -17.684753 *and* Ankle.ACC.Mean $>$ 5.747455 **then fatigued** ($C_a = 0.094675$)
9) **if** Chest.ACC.Mean $\leq$ 2.328720 **then fatigued** ($C_a = 0.041420$)

Comparing the two rulesets, first it is possible to observe that the augmentation process increases the number of generated rules (being 5 rules for real data only and 9 with real+fake). The higher similarity is mostly related to features that are shared in both rulesets: in detail, *Hip.yposture.Mean* and *Hip.ACC.Mean*.

VIII. Conclusions

In this paper, we propose an explainable method, based on the LLM algorithm, to understand and evaluate GAN-based data augmentation and its advantages. Our approach extends the typical usage of GANs, in which the data augmentation is applied to outperform baseline results.

We experiment our scheme in the context of physical fatigue detection, showing how GANs can be adopted to generate two subgroups of the original dataset, in our case based on the age of the subjects, up to 40 or over 40 years old. After performing different runs of GANs, we compare the obtained datasets through LLM accuracy and F1-score in different scenarios; then, an innovative measure of rule similarity is used to find the best among different generated fake datasets and discovering new knowledge.

Future works on the topic may include improvements in GANs training phase, to reach better parameters, or improvements to the architecture to fasten the process. Also, in this paper, we considered scenario 2 (combination of fake and real data) as the best option: then, another future direction will be to investigate more on scenarios 1 and 3. Moreover, future developments may involve the testing of different fake/real sizes proportions, in order to assess how the system scales with the size of the data.

ACKNOWLEDGEMENTS

CNR-IEIIT was supported by Fondazione Compagnia di San Paolo, scientific research call 2019 (Bando 2019-2020 per progetti di ricerca scientifica presentati da enti genovesi): project "Advances in pneumology via ICT and data analytics" (PNEULYTICS), Prot. No 2020.AAI22.U41/SD/pv; practice number: 2019.0988, ID ROL: 34754. This work was also supported by Robotics and AI for Socio-economic Empowerment (RAISE) project, Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR), ecosystem innovation call 2022, Spoke 2 - Smart Devices and Technologies for Personal and Remote Healthcare.

REFERENCES

[1] K. Naseer Qureshi, S. Din, G. Jeon, and F. Piccialli, "An accurate and dynamic predictive model for a smart m-health system using machine learning," *Information Sciences*, vol. 538, pp. 486–502, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520306113

[2] A. Adadi and M. Berrada, "Explainable ai for healthcare: From black box to interpretable models," in *Embedded Systems and Artificial Intelligence*. Springer, 2020, pp. 327–337.

[3] I. Ahmed, G. Jeon, and F. Piccialli, "A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 318–10 326, 2021.

[4] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Incorporating explainable artificial intelligence (xai) to aid the understanding of machine learning in the healthcare domain," in *Proc. of the 28th Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, 12 2020.

[5] A. B. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[6] M. M. Rahman and D. N. Davis, *Machine Learning-Based Missing Value Imputation Method for Clinical Datasets*. Dordrecht: Springer Netherlands, 2013, pp. 245–257.

[7] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of healthcare engineering*, vol. 2019, 2019.

[8] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 979.

[9] Q. Jin, H. Cui, C. Sun, Z. Meng, and R. Su, "Free-form tumor synthesis in computed tomography images via richer generative adversarial network," *Knowledge-Based Systems*, vol. 218, p. 106753, 2021.

[10] M. Frid-Adar *et al.*, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 289–293.

[11] C. Han *et al.*, *Infinite Brain MR Images: PGGAN-Based Data Augmentation for Tumor Detection*. Singapore: Springer Singapore, 2020, ch. 27, pp. 291–303.

[12] V. Ramesh, K. Vatanparvar, E. Nemati, V. Nathan, M. M. Rahman, and J. Kuang, "Coughgan: Generating synthetic coughs that improve respiratory disease classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5682–5688.

[13] S. Haradal, H. Hayashi, and S. Uchida, "Biosignal data augmentation based on generative adversarial networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 368–371.

[14] J. Geogres-Filteau and E. Cirillo, "Synthetic observational health data with gans: from slow adoption to a boom in medical research and ultimately digital twins?" https://arxiv.org/pdf/2005.13510.pdf, 2020.

[15] A. Torfi and E. A. Fox, "Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records," in *The Thirty-Third International Flairs Conference*, 2020, pp. 335–340.

[16] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Machine learning for healthcare conference*. PMLR, 2017, pp. 286–305.

[17] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, and M. Mongelli, "A generative adversarial network (gan) technique for internet of medical things data," *Sensors*, vol. 21, no. 11, 2021.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[19] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[21] Rulex analytics platform, https://www.rulex.ai/.

[22] M. Muselli, "Switching neural networks: A new connectionist model for classification," pp. 23–30, 2005.

[23] M. Mongelli, "Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence," *Computer Communications*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366421002504

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

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

[26] Z. Sedighi Maman, Y.-J. Chen, A. Baghdadi, S. Lombardo, L. A. Cavuoto, and F. M. Megahed, "A data analytic framework for physical fatigue management using wearable sensors," *Expert Systems with Applications*, vol. 155, p. 113405, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420302293

[27] N. Williams, "The Borg Rating of Perceived Exertion (RPE) scale," *Occupational Medicine*, vol. 67, no. 5, pp. 404–405, 07 2017. [Online]. Available: https://doi.org/10.1093/occmed/kqx063