Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models

Marta Lenatti*, Alessia Paglialonga*, Vanessa Orani*, Melissa Ferretti*, and Maurizio Mongelli*

***Abstract*— The aim of this study is to apply and characterize eXplainable AI (XAI) to assess the quality of synthetic health data generated using a data augmentation algorithm. In this exploratory study, several synthetic datasets are generated using various configurations of a conditional Generative Adversarial Network (GAN) from a set of 156 observations related to adult hearing screening. A rule-based native XAI algorithm, the Logic Learning Machine, is used in combination with conventional utility metrics. The classification performance in different conditions is assessed: models trained and tested on synthetic data, models trained on synthetic data and tested on real data, and models trained on real data and tested on synthetic data. The rules extracted from real and synthetic data are then compared using a rule similarity metric. The results indicate that XAI may be used to assess the quality of synthetic data by (i) the analysis of classification performance and (ii) the analysis of the rules extracted on real and synthetic data (number, covering, structure, cut-off values, and similarity). These results suggest that XAI can be used in an original way to assess synthetic health data and extract knowledge about the mechanisms underlying the generated data.**

***Index Terms*— data augmentation, eXplainable AI (XAI), hearing screening, rule similarity, Generative Adversarial Networks (GAN).**

I. Introduction

The area of synthetic data generation is gaining growing attention in healthcare. Generation of high-quality synthetic data can help build realistic datasets that can be shared openly in the educational and scientific community, for example to support the development of predictive models of disease, averting issues related to patient identification and data privacy that frequently limit the widespread use of health data [1]–[4]. Data augmentation from patient monitoring devices can help limit issues related to missing data, misuse, or lack of compliance [5]–[7]. Synthetic data generation can help develop large datasets from small ones as well as balanced datasets from highly unbalanced ones, and it can help limit the costs of building datasets from large cohorts of patients [8], [9]. The goal of data augmentation algorithms is to create realistic and useful synthetic data, namely preserving distributions, predictive capabilities, and relationships [1], [9]. The field of data generation algorithms is still an important area of research, however it is beyond the aims of this study to develop and assess synthetic data generation techniques. Rather, this study focused on introducing and characterizing novel metrics to assess the quality of synthetic data. Several approaches have been introduced in the literature, such as utility metrics derived from the distributions (e.g., Maximum Mean Discrepancy (MMD), Hellinger distance, Classifier Two Sample (C2S) metric), or measures based on classification performance on real and generated data [10]–[14].

Utility metrics and classification performance can give a general picture of the quality of generated data but they provide limited insight into the way input-output relationships are preserved in synthetic data. EXplainable AI (XAI) techniques could help assess if, and to what extent synthetic data maintain input-output relationships similar to those found in real data [15], [16]. When dealing with health data, XAI methods are particularly promising as they can help healthcare experts enter the logic of the machine learning process and extract knowledge about the mechanisms underlying the observed phenomena in a meaningful and transparent way, so that synthetic data can be validated against available knowledge [17], [18]. In a preliminary study on data from a pilot experiment on respiratory disease monitoring, we showed that conventional utility metrics are able to anticipate XAI classification performance, being low utility metrics associated with low classification performance of XAI models trained on synthetic data and tested on real data [6]. However, the ability of XAI to provide additional information about the logic underlying synthetic health data has not been specifically investigated so far. The aim of this study is to apply and characterize XAI-based models and metrics as a means to assess the quality of synthetic health data. The novel contributions here introduced consist of: XAI evaluation of synthetic datasets in terms of feature relevance, visual inspection of rules, and classification performance; and the definition of a new rule similarity method to compare rule-based XAI models trained on synthetic and real data.

II. Dataset

The example of health dataset assessed in this study includes hearing screening data collected from a self-administered adaptive speech-in-noise test for adult hearing screening in the context of project WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk) [19], [20]. Multivariate hearing screening data are particularly useful as they can be used to develop machine learning models able to identify individuals with hearing loss, therefore supporting widespread screening of this largely underdiagnosed condition, that is currently the third leading cause of years lived with disability worldwide [21], [22]. However, there is scarcity of multivariate hearing screening data to build machine learning models to predict hearing loss as, to date, screening outcomes are typically determined on the basis of a single variable (e.g., the speech recognition threshold, SRT, or the number/percentage of correct responses) [23].

The WHISPER dataset includes 156 records related to eight input features extracted from speech-in-noise testing and one output class, that is the presence or absence of hearing loss in the tested ear, as determined by the pure tone average (PTA), i.e. the average value of pure-tone thresholds measured at 0.5, 1, 2, and 4 kHz. The output class is defined following the World Health Organization (WHO) definition of slight/mild hearing impairment, in force until Feb 28, 2021 [24]: hearing loss (“HL”, PTA $>$ 25 dB HL: 55 records) and no hearing loss (“no HL”, PTA $\leq$ 25 dB HL; 101 records). The input features extracted upon completion of the speech-in-noise test comprise: subject’s *age*, *SRT*, measured in dB signal-to-noise ratio; *#trials*, i.e. number of presented stimuli; *#correct*, i.e. number of

Article submitted on Jan 31, 2022. The work leading to the creation of the WHISPER dataset was partially supported by Capita Foundation (project WHISPER, Widespread Hearing Impairment Screening and PrEvention of Risk, 2020 Auditory Research Grant).

All authors are with the National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT) (e-mail:marta.lenatti@ieiit.cnr.it; corresponding author A. Paglialonga: phone: +39-02-23993343, e-mail: alessia.paglialonga@cnr.it; vanessa.orani@ieiit.cnr.it; melissa.ferretti@ieiit.cnr.it; maurizio.mongelli@ieiit.cnr.it.

* all authors equal contribution.

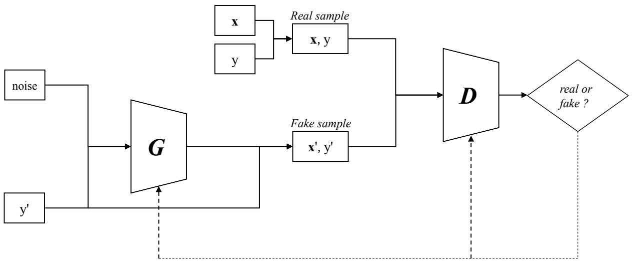

Fig. 1. Scheme of a conditional GAN describing an adversarial learning game between a generator G and a discriminator D.

correct responses; *%correct*, i.e. percentage of correct responses; *avg reaction time*, i.e. average time needed to provide a response; *total test time*, i.e. total time needed to complete the test; and *volume*, i.e. self-adjusted volume set by the participant before taking the test, computed on a range from 0 to 1. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion n. 2/2019, Feb 19 2019).

III. Generation of synthetic data

In this study, synthetic data are generated using Generative Adversarial Networks (GAN) [25], a deep learning approach able to reach remarkable performance in generating high-quality synthetic data, for example in the field of images [26], biosignals [27], [28], and time series from patient monitoring devices [6]. A GAN comprises two neural networks: a generator (G) for generating fake but realistic data x', and a discriminator (D) for distinguishing whether the generated data are real or fake. Learning is achieved by an adversarial game between G and D: G uses the encoder-decoder scheme to build synthetic data, whereas D infers the separation between real and synthetic data. Therefore, D learns to become better at distinguishing real from synthetic data and G learns to generate better data to fool the discriminator [25]. In this study, a conditional GAN [29], [30] is implemented (see Fig. 1), i.e. a GAN in which G and D are conditioned during training by using output class labels in a way that G learns to produce realistic examples for each label in the training set starting from random noise, and D learns to distinguish fake example-label pairs $(\mathbf{x}', y')$ from real example-label pairs $(\mathbf{x}, y)$. A set of balanced synthetic datasets are generated by varying different GAN parameters, namely the number of nodes per layer in G and D networks, the batch size, and the number of epochs, as follows:

1) G nodes: 128, 64, 32, 1; D nodes: 32, 64, 128; Batch size: 64.
2) G nodes: 64, 32, 16, 1; D nodes: 16, 32, 64; Batch size: 32.
3) G nodes: 64, 32, 16, 1; D nodes: 16, 32, 64; Batch size: 64.

For each of these three configurations, the number of epochs is set at five different values: 10000, 15000, 20000, 25000, and 30000, to obtain a total of 15 different synthetic datasets.

IV. Assessment of synthetic data using utility metrics

To monitor the quality of the GAN generation process, we use a combination of the following measures: MMD, C2S metric, Hellinger Distance (HD) and Pairwise Correlation Difference (PCD) [10].

The MMD metric is a measure of dissimilarity between two probability distributions P and Q that uses samples drawn independently from each of them [10]. Given a kernel k and its associated Reproducing Kernel Hilbert Space (RKHS) H_k of functions defined on a set X, the distance between the two probability distributions P and Q in the original space is converted into a distance between their relative mean embeddings of features in the space H_k [31]. A statistical hypothesis test is introduced to test the null hypothesis $H_0 : P = Q$ versus the alternate hypothesis $H_1 : P \neq Q$ [31]. The test statistic is compared to a threshold which depends on the probability P and K and is selected based on the chosen α level. In this study, a Gaussian radial basis function kernel (rbf) is chosen for the MMD statistical test.

The C2S metric uses a machine learning classifier to assess whether two samples are drawn from the same distribution [10], [32], [33]. The C2S metric computation comprises the following steps:

1) A dataset D is built by combining the real samples as 0 and the synthetic samples as 1;
2) The dataset is randomly split into two disjoint training and testing subsets (D_{train} and D_{test}, respectively);
3) A binary classifier (e.g., logistic regression) is trained on D_{train} and the C2S metric is defined as the classification accuracy of this classifier computed on D_{test}.

Hence, the higher the C2S metric (i.e., the accuracy), the more likely the two distributions are different, whereas for samples drawn from the same distribution the accuracy should remain near chance-level. To maintain class balance between real and generated data, the MMD and C2S metrics are computed using the same number of samples as in the original dataset and then averaged across 10 random realizations of the sampled subsets.

The HD [14] is a utility metric related to the Bhattacharyya coefficient-based measure that evaluates the distance between two probability distributions in their original space. This metrics ranges from 0 (i.e., identical distributions) to 1 (i.e., totally dissimilar distributions). The HD has been derived in this study starting from the probability density functions of the datasets to be compared.

Finally, PCD [12] was evaluated to investigate if synthetic datasets were able to retain the correlations among features that characterize the original distribution. The PCD between a real and a synthetic dataset is defined as the Frobenius norm of the difference of the correlation matrices extracted from the two datasets to be compared. The lower the PCD, the greater the similarity between the correlations in the original dataset and those in the synthetic dataset.

V. Assessment of synthetic data using XAI

Among the various XAI techniques available, in this study we used the Logic Learning Machine (LLM), a technique able to generate transparent models whose inner logic could be described using a set of n intelligible rules, in the form *if (premise) then (consequence)*, where premise is a logical product of m conditions c_j, and consequence provides a class assignment for the output y [34], [35][1]. Let $x_1, ..., x_n$ be the input features, each defined in a specific domain. Then, a condition involving the variable x_j, can have one of the following forms: $x_j > \lambda, x_j \leq \mu, \lambda < x_j \leq \mu$, being λ, μ values belonging to the feature's domain. The classification uncertainty associated with a rule $R(i)$ is described by the measures of covering and error shown in (1) and (2):

$$Covering : C(R(i)) = \frac{TP(R(i))}{TP(R(i)) + FN(R(i))} \tag{1}$$

$$Error : E(R(i)) = \frac{FP(R(i))}{FP(R(i)) + TN(R(i))} \tag{2}$$

where $TP(R(i)), FP(R(i)), TN(R(i))$, and $FN(R(i))$ are the true positives, false positives, true negatives, and false negatives associated

[1] The Rulex platform, www.rulex.ai, is used as it contains a big data implementation of the LLM.

with the rule $R(i)$. Feature relevance is derived from (1) and (2). In order to obtain the relevance $Rel(c_j)$ of a condition, we compare the rule R, in which condition c_j occurs, and the same rule without that condition, called R'. Since the premise part of R' is less stringent, we obtain that $E(R') \geq E(R)$, thus the quantity $Rel(c_j) = (E(R') - E(R))C(R)$ indicates the relevance for the condition of interest and, therefore, for the feature involved in that condition.

A. Analysis of classification performance

The real and synthetic datasets were randomly split into training and test sets by applying stratification. The classification performance was addressed by computing sensitivity, specificity, and F1-score in LLM models deployed with the following combinations of training (Tr) and test (Te) sets:

- Condition A (baseline): TrR = training set from real dataset (80%), TeR = test set from real dataset (20%)
- Condition B: TrS = training set from synthetic dataset (80%), TeS = test set from synthetic dataset (20%);
- Condition C: TrS = training set from synthetic dataset (80%), TeR = test set is the whole real dataset ;
- Condition D: TrR = training set from real dataset (80%), TeS = test set is the whole synthetic dataset.

A cross-classification (CC) measure [12] was introduced to summarize the similarity between real and synthetic datasets in terms of classification performance. Two CCs were computed as the ratio of the accuracy in conditions C and D to the accuracy in condition A.

B. Analysis of similarity between rules

A measure of similarity between rules is introduced, based on the cosine similarity between Bag of Words (BOW) [36] representations of the set of rules extracted from real and synthetic datasets. BOW is a widely used text representation approach (e.g., [37], [38]) where a text is decomposed into a matrix of words and their relative frequencies. In a preliminary study [39], a BOW-based metric was introduced to individually compare rules from different classes of the same dataset and rule sets referring to stratifications (e.g., different age groups) of the same phenomenon. In this study, we further elaborated this metric by considering the difference in covering between different rules and introducing a *global similarity metric* that, based on the similarity between pairs of rules, provides an estimate of similarity between rule sets, i.e., between the models that describe the underlying data. Each rule $R(i)$, associated to the output class y, can be defined by a set of m conditions, each described by a word w (i.e., the combination of the feature name and direction of the inequality sign) and the related cut-off value t as shown in [35].

$$R(i) = \text{if } \{c_j\}_{j=1}^{m} \text{ then } y, \qquad y \in [0,1] \tag{3}$$

$$c_j = (w_j(i), t(i)) \tag{4}$$

Two rules can be considered similar when their conditions share the same structure (i.e., same feature and same direction) and similar cut-off values [40]. In the specific case of classification rules, there can be at maximum one condition for each feature (i.e., a word can be present only once in the rule), so the related cells of the BOW matrix contain binary values (1 if the word is present and 0 if the word is not present). For each word, an additional column is added to account for the cut-off value, normalized between 0 and 1 based on the theoretical lowest and highest possible values of the feature.

Once the BOW matrix is created for both rulesets to be compared, cosine similarity is applied to all the combination of couples of rules $\left\{R_{real}(i_r)_{i_r=1}^{m_r}, R_{synthetic}(i_s)_{i_s=1}^{m_s}\right\}$, divided by class, to obtain a measure of similarity between rules S'_{rs}. Cosine similarity is a widely used text similarity measure, often combined with BOW representation (e.g., [41]), that measures the similarity between two vectors in terms of the cosine of the angle in between. To compute rule similarity, only rules with covering higher than 15% are considered, as rules with lower covering are representative of only a few input data and therefore may be subject to greater variability due to the choice of training and test partitions, especially in small datasets like the one here used. Intuitively, if the real and synthetic datasets are similar, their rules should be similar in terms of structure and covering. Hence, the difference in covering between rules extracted from real and synthetic data is introduced as a weighting factor in the computation of rule similarity. Therefore, the resulting similarity metric is:

$$S_{rs} = S'_{rs} * \left(1 - \left|C(R_{real}(i_r)) - C(R_{synthetic}(i_s))\right|\right) \tag{5}$$

where $C(R_{real}(i_r))$ and $C(R_{synthetic}(i_s)$ are the covering of the real rule and of the synthetic one, respectively[2] A global similarity metric between rulesets G_x is defined as the ratio of the number of real-synthetic rule pairs n_x with similarity greater than a predetermined threshold value x (i.e, 0.6 in this study) to the total number of rules extracted from the real dataset m_r.

VI. Results

A. WHISPER dataset

Hold-out cross validation of LLM models using ten randomly shuffled versions of the training and test set was performed to extract feature relevance (as defined in SectionV). The most relevant features are: age, #correct, SRT, %correct, and avg reaction time, (relevance: age = 0.67 ± 0.07; #correct = 0.52 ± 0.21; SRT = 0.40 ± 0.22; %correct= 0.22 ± 0.19; avg reaction time= 0.21 ± 0.17). Vice versa, test volume, total test time, and number of trials do not contribute substantially to the output class (relevance: volume = 0.03 ±0.03; total test time = 0.10 ±0.07; #trials: 0.15 ±0.24).

For the sake of simplicity, a reduced version of the experimental dataset including the three most relevant features (age, #correct, SRT), and the output class is here used to assess the outcomes of XAI on synthetic data and enable straightforward visualization and interpretation of results. Table I shows the MMD, the related p-value, the HD and the C2S metric as a function of the GAN settings for the 15 synthetically generated datasets. For the MMD and C2S metrics, the mean and standard deviation (s.d.) are computed over 10 iterations as described in Section III. The results in Table I suggest that the synthetic datasets more similar to the real one in terms of MMD (lower values, p-value $>$ 0.05), HD (lower values e.g., $<$ 0.40), and C2S metrics (near chance level), are #8, #9, #13, and #15, however, on the basis of the observed metrics, no straightforward indication of the 'most similar' synthetic dataset can be derived. The PCD was calculated to assess whether the synthetic datasets are able to maintain correlations between features that resemble those in the original dataset. PCD values obtained for the datasets with better MMD, C2S, and HD are very close to each other (i.e., $PCD_{\#8} = 1.01$, $PCD_{\#9} = 0.83$, $PCD_{\#13} = 0.65$, $PCD_{\#15} = 0.81$). Moreover, PCD values are better (i.e., smaller) in the datasets mentioned above than in the synthetic datasets with worse values of MMD, C2S, and HD (e.g., $PCD_{\#11} = 2.51$).

1) Analysis of classification performance: The LLM model trained on the WHISPER dataset includes 12 rules overall (7 for "no HL", average covering = 25.32%; 5 for "HL", average covering = 25.65%). The rule with highest covering for class "no HL"

[2] The Python code and exemplary rulesets are available open source: `https://github.com/lenattimarta/BOW_rule_similarity`.

TABLE I

MMD, HD AND C2S METRICS FOR THE SYNTHETICALLY GENERATED DATASETS AS A FUNCTION OF THE GAN SETTINGS.

#	#epochs	MMD metric		HD	C2S metric
		$mean \pm sd(x10^-2)$	p-value	$mean$	$mean \pm sd$
1) G units: 128,64,32,1; D units:32,64,128; Batch size: 64					
1	10000	6.33 ± 0.10	0.003	0.50	0.72 ± 0.05
2	15000	6.11 ± 0.10	0.033	0.46	0.69 ± 0.04
3	20000	5.52 ± 0.03	0.042	0.46	0.61 ± 0.05
4	25000	5.55 ± 0.10	0.095	0.40	0.60 ± 0.06
5	30000	9.50 ± 0.30	0.000	0.67	0.91 ± 0.02
2) G units: 64,32,16,1; D units: 16, 32, 64; Batch size: 32					
6	10000	5.94 ± 0.10	0.177	0.46	0.75 ± 0.04
7	15000	6.32 ± 0.20	0.049	0.46	0.68 ± 0.07
8	20000	5.21 ± 0.10	0.950	0.38	0.51 ± 0.06
9	25000	5.40 ± 0.10	0.776	0.39	0.49 ± 0.04
10	30000	5.62 ± 0.10	0.254	0.41	0.62 ± 0.09
3) G units: 64,32,16,1; D units: 16, 32, 64; Batch size: 64					
11	10000	6.77 ± 0.20	0.002	0.47	0.71 ± 0.07
12	15000	5.55 ± 0.10	0.116	0.40	0.59 ± 0.10
13	20000	5.15 ± 0.10	0.995	0.37	0.48 ± 0.04
14	25000	5.49 ± 0.10	0.155	0.40	0.55 ± 0.05
15	30000	5.28 ± 0.10	0.369	0.39	0.51 ± 0.03

$(R_{r,noHL}1)$ indicates that subjects younger than 52 years are more likely to have better hearing ability than older subjects, in line with the well-known relationship between age and hearing loss [42]. The second rule with highest covering for class “no HL” $(R_{r,noHL}2)$ indicates that subjects with a negative SRT (i.e., below -7.35 dB SNR) who achieve good results in the speech-in-noise test (i.e., more than 96 stimuli correctly identified) will probably belong to the normal hearing class. This rule synthesizes well the relationship between speech recognition ability and hearing loss. Conversely, subjects with a poor performance of speech recognition in noise (i.e., lower than 59 correct responses) as in $R_{r,HL}1$ will more likely suffer from hearing loss [23], [43].

Fig. 2 shows the classification performance on the test set (sensitivity, specificity, and F1-score) of the four synthetic datasets with low MMD and HD (#8, #9, #13, and #15) and a synthetic dataset with high MMD and HD (#11), as computed in the conditions A, B, C, and D defined in Section V-A. The degree of overfitting in the analyzed models was assessed by evaluating the difference between training and test accuracy of the LLM models obtained with the four different combinations of training and test set. The following mean differences were calculated for the datasets considered: condition A = 5.72% (sd=3.6%) (average difference in performance obtained with 5-fold-cross validation), condition B = 2.28% (sd=1.1%), condition C = 11.17% (sd=1.8%), condition D = 12.43% (sd=4.8%). The discrepancy between training and test performance is limited (i.e., lower than 13% on average), including conditions C and D in which training and test portions are extracted from different datasets. Overall, the test performance is satisfactory, with accuracy around 75%-80% in the models with lower classification performance, thus demonstrating limited overfitting.

Generally, the performance metrics measured on the synthetic datasets #8, #9, #13, and #15 are higher than those measured on datasets #11, reflecting the well-known capability of the MMD to discriminate between datasets that are significantly different from the original one and datasets that are similar to the original one. In condition B, the classification performance of models trained and tested on synthetic datasets #8, #9, #13, and #15 is similar to or higher than that of models trained and tested on real data (condition A). Specifically, higher specificity and F1-score, and similar sensitivity is observed. In condition C, i.e. the condition in which the capability of synthetic models to be applied on real data is assessed, synthetic models from datasets #8, #9, #13, and #15 maintain a similar specificity and F1-score with respect to real data, but are characterized by a slightly lower sensitivity, suggesting that models trained on synthetic datasets are in general less able to detect the ‘HL’ class, when applied on real data, compared to real models. In condition D, i.e. the condition in which the capability of the real model to classify synthetic data is evaluated, similar F1-score, higher specificity, and a drop in sensitivity are observed compared to condition A. The cross-classification based on test accuracy in condition C yields the following results: $CC_{\#8} = 0.98$, $CC_{\#9} = 0.97$, $CC_{\#13} = 0.94$,

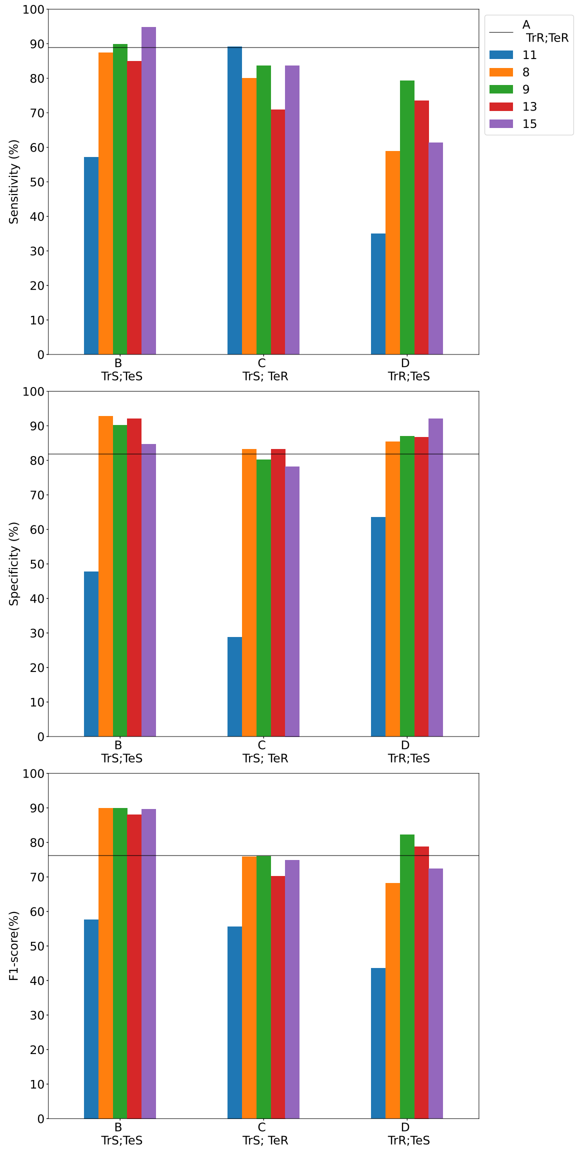

Fig. 2. Test classification performance in terms of sensitivity (top panel), specificity (center panel), and F1-score (bottom panel) of five synthetic datasets (#11: p-value MMD test $<$0.05; #8, #9, #13, #15: p-value MMD test $\geq$ 0.05), with respect to the real classification performance.

$CC_{\#15} = 0.96$, $CC_{\#11} = 0.60$. The cross-classification based on the test accuracy in condition D yields the following results: $CC_{\#8} = 0.86$, $CC_{\#9} = 0.99$, $CC_{\#13} = 0.95$, $CC_{\#15} = 0.92$, $CC_{\#11} = 0.56$. The classification performance is similar to the real one (i.e., CC close to 1) for synthetic datasets with lower MMD, HD and C2S, whereas classification performance is worse (i.e., CC lower than 1) for the synthetic dataset with higher MMD, HD and C2S metrics.

2) Analysis of similarity between rules: Table II shows the rule similarity coefficients, as defined in (5), obtained by comparing the LLM model trained on the real dataset with those trained on the synthetic datasets #8, #9, #13, and #15, i.e. the ones that are not significantly different from the real dataset, according to the MMD, HD, and C2S metrics. A global metric of comparison between rulesets G_x is shown in the last column, defined as the ratio of the number of real-synthetic rule pairs with similarity greater than 0.6 to the total number of rules extracted from the real dataset. For the sake of clarity, only rules with covering higher than 15% are considered. The rules are reported in full detail in **Appendix I**.

For most of the rules extracted from the real dataset there is at least one rule with similarity greater than 0.3 in each of the four synthetic datasets considered. It is worth noting that the rule similarity measure here used considers the rule structure, the cut-off values and the related covering, as defined in Section V-B. For example, from each of the four synthetic datasets a rule in the form $Age \leq \mu_{Age}$ is observed, that is very similar to the one extracted from the real dataset ($R_{r,noHL}1$: $Age \leq 52$), but the resulting similarities are slightly different, mainly due to differences in covering. The highest value of rule similarity has been identified for the rule $R_{15,noHL}5$ that is similar to the real rule $R_{r,noHL}3$ ($SRT \leq -16.11$; C: 21.74%) in terms of both structure and covering ($SRT \leq -17.75$; C: 26.37%). Among the four synthetic datasets here assessed, #15 is the one with the highest global similarity G_x.

Fig. 3 shows a visual overview of the rules extracted from the real dataset, from the optimal dataset, i.e. the one with low MMD, HD and C2S metrics *and* high rule similarity (#15), from a dataset with low MMD, HD and C2S metrics but relatively low global similarity (#9) and from a dataset with high MMD, HD and C2S metrics (#11). The inner circular crowns represent the rules of each model in terms of covering (outer diameter), error (inner diameter), and class (color) whereas the outer slices represent the values of each of the three input features in terms of class (color) and relevance (opacity). The rules extracted from the synthetic datasets #15 and #9 are more similar to the ones obtained from the real dataset in terms of number, covering, and error compared to those extracted from the synthetic dataset #11, that is associated with a higher number of rules, lower covering, and higher error. In terms of value ranges associated with the two output classes, as shown in the outer slices, the synthetic dataset #15 shows a clear separation of the two classes for each of the three input features (cut-off values: Age: 49 years; #correct: 65; SRT: -9.49 dB SNR), with cut-off values that are similar to those observed in the real dataset (Age: 52 years; #correct: 64; SRT: -10.16 dB SNR). The model trained on dataset #9 presents similar cut-off in terms of #correct (i.e., 63), a clear, but higher, cut-off on age (i.e., 66), but no clearly defined cut-off on SRT. Conversely, no clearly identifiable cut-off values are found in the model trained on dataset #11 as the features are distributed in a similar way between the two classes.

VII. Discussion

Synthetic data generation may be of help in creating large, balanced, de-identified medical datasets that can be used to train and validate new AI algorithms to improve disease detection and prediction, overcoming common problems in real-world clinical datasets such as data scarcity and class imbalance [5]–[8]. Trustworthiness of medical decisions supported by AI models becomes essential, especially when the model has been built using synthetic or augmented data [44]. In this context, XAI techniques may enable transparent data generation and analysis, allowing the end user to understand the logic of the model and decide whether to trust and validate its decisions.

In this exploratory study, we propose and characterize a framework of XAI as a means to assess the quality of synthetic tabular data. Specifically, a fully interpretable algorithm (the LLM) is used to generate rule-based models of the data in order to simultaneously assess distributions, predictive capabilities, and relationships in synthetic data by the analysis of the set of rules and the related classification performances.

For a first characterization of the proposed approach, a dataset including multivariate measures of hearing performance, with a single record per subject (dataset WHISPER, 156 records) is considered. This dataset was chosen as an example, but the proposed approach is general and can be extended to different applications. Synthetic data (1000 records) are generated from this real dataset by using a conditional GAN and by systematically varying the number of G and D nodes, the batch size, and the number of epochs.

A. Assessment of synthetic data using XAI

Datasets with significantly different values of MMD, HD, and C2S metrics are characterized by different levels of quality. Vice versa, when dealing with different synthetic datasets that exhibit similar values of utility metrics such as the MMD, HD and C2S metrics here used, quantitative analysis of XAI in terms of classification performances and inspection of decision rules is helpful to assess the similarity between synthetic and real data.

An example of application based on the WHISPER dataset including a subset of the most relevant input features (i.e., SRT, age, #correct), and output class is proposed in Section VI-A. Specifically, four different datasets similar to the real dataset based on MMD (i.e., low MMD, from $5.13x10^{-2}$ to $5.40x10^{-2}$, p-value $>$ 0.05) and HD values (i.e., low HD, $<$ 0.40) are compared in terms of classification performance (Fig. 2). LLM models trained on the selected synthetic datasets (#8, #9, #13, and #15) have, on average, slightly lower sensitivity with respect to the LLM model trained on real data, when tested on real data (condition C), thus they are generally less able to detect the target output class. Vice versa, synthetic models are on average better in identifying normal hearing subjects (i.e., higher specificity). All the LLM models trained on the selected synthetic datasets maintain a satisfactory classification performance, remarkably similar to the performance of the model trained on real data, as demonstrated by the cross-classification metric.

B. Analysis of similarity between rules

In this study a rule similarity metric (5), defined as a combination of similarity in rule structure, cut-off values, and covering, is introduced to assess possible differences between the sets of rules that characterize the models extracted from different synthetic datasets. Rule similarity analysis highlights that the LLM model trained on synthetic dataset #15 is described by rules that are closer to those of the real model (i.e., higher G_x) , with respect to those of the other candidate datasets (Table II). Rule visualization (Fig. 3) helps intuitively appreciate the differences in LLM models trained on the real dataset, the optimal synthetic dataset (#15 i.e., low MMD, HD and C2S metrics, highest G_x), a suboptimal synthetic dataset (#9 i.e., low MMD, HD and C2S metrics, low G_x) and an example where data generation process has not achieved the desired results (#11 i.e.,

TABLE II

RULE SIMILARITY COEFFICIENTS OBTAINED COMPARING THE REAL DATASET WITH THE SYNTHETIC DATASETS #8, #9, #13, AND #15.

Real Synthetic		$R_{r,noHL}1$	$R_{r,noHL}2$	$R_{r,noHL}3$		$R_{r,HL}1$	$R_{r,HL}2$	$R_{r,HL}3$	$R_{r,HL}4$	$G_{0.6}$
#8	$R_{8,noHL}1$	0.92	0	0	$R_{8,HL}1$	0.38	0.3	0.37	0.20	3\7
	$R_{8,noHL}2$	0.44	0.66	0.32	$R_{8,HL}2$	0.7	0.37	0.43	0.29	
#9	$R_{9,noHL}1$	0.64	0.25	0	$R_{9,HL}1$	0.5	0.25	0.25	0.16	1\7
	$R_{9,noHL}2$	0.93	0	0	$R_{9,HL}2$	0	0.46	0.25	0.34	
	$R_{9,noHL}3$	0.43	0.43	0.42						
#13	$R_{13,noHL}1$	0	0.84	0.41	$R_{13,HL}1$	0	0.56	0.31	0.41	2\7
	$R_{13,noHL}2$	0.90	0	0	$R_{13,HL}2$	0.59	0.32	0.34	0.23	
	$R_{13,noHL}3$	0.26	0.57	0.33	$R_{13,HL}3$	0	0.58	0.39	0.54	
	$R_{13,noHL}4$	0.28	0.44	0.38						
#15	$R_{15,noHL}1$	0	0.64	0.50	$R_{15,HL}1$	0	0.39	0.37	0.28	5\7
	$R_{15,noHL}2$	0.87	0	0	$R_{15,HL}2$	0.37	0.72	0.82	0.58	
	$R_{15,noHL}3$	0.41	0.62	0.45						
	$R_{15,noHL}4$	0	0.61	0.47						
	$R_{15,noHL}5$	0	0.36	0.96						

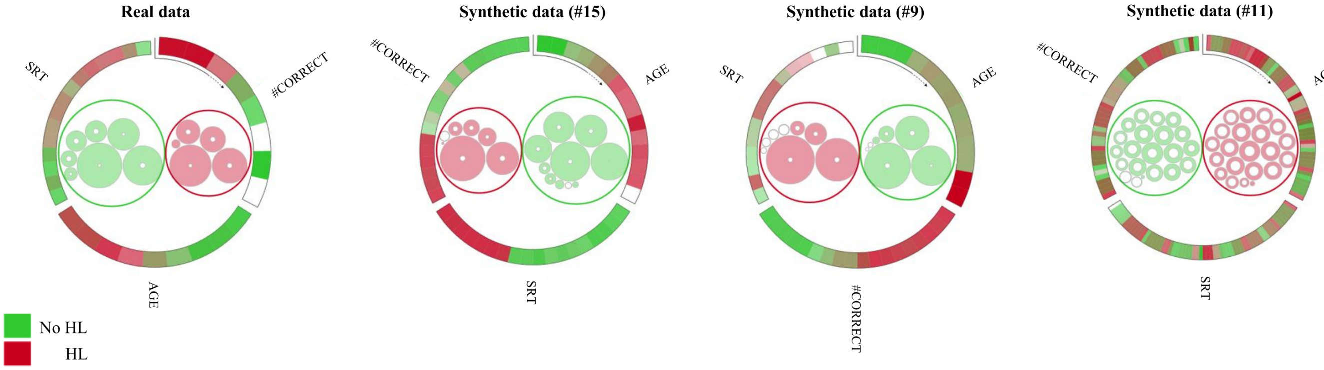

Fig. 3. Rule visualization for the real dataset and for the synthetic datasets #15, #9, and #11.

high MMD, HD and C2S metrics). As it can be noticed in Fig. 3, the inner logic of the LLM model trained on dataset #15 resembles that of the real one, by maintaining similar input-output relationships and cut-off values. Moreover, the data augmentation process seems to simplify the intrinsic behavior of certain variables, by cleaning up some regions of uncertainty in classification. For example, the model trained on synthetic dataset #15 amplifies the well-known relationship between SRT and hearing loss and allows us to define a cut-off at -9.49 dB SNR which is similar to the one suggested by previous studies (e.g., [43]). As expected, the LLM model trained on the synthetic dataset #11 (worse MMD and C2S metrics) has a much higher number of rules, with lower average covering, different structure and different cut-off values, than the one trained on the real dataset. Rule similarity analysis provides additional information about the quality of the datasets compared to statistical measures derived from distributions (e.g., utility metrics like MMD, HD, and C2S metrics) or from model testing (e.g., classification performance in conditions B, C, and D). The synthetic datasets that pass the MMD, HD, and C2S tests (#8, #9, #13, #15) are then filtered by rule similarity, that confirms their quality as they all present one or more rules with similarity higher than 0.3 when compared to the real rules. However, for some of these datasets (e.g., dataset #15) higher G_x is observed, suggesting a higher similarity to the real dataset in terms of input-output relationships. Therefore, the proposed rule similarity metric allows us to select a specific dataset, within a set of good-quality datasets that are considered equally similar in terms of utility metrics. For the computation of a global metric of comparison between datasets, in this preliminary study rule similarity has been considered as high when it exceeds 0.6, however this value needs to be further validated. The proposed metric has been applied to LLM models, but it is in principle applicable to other native rule-based methods (e.g., Decision Trees) or black box models made explainable by post-hoc XAI methods. For example, visual inspection of partial dependence plots estimated from Random Forest models trained on real and synthetic datasets shows that the averaged partial dependence trends obtained from the synthetic datasets #15 and #9 are similar to the one obtained from the real dataset, and their approximate cut-off values are similar to the cut-off values of the rules as shown in Fig. 3. However, for partial dependence plots and, more generally, for post-hoc XAI techniques, further processing is needed to determine decision rules and further research in this direction would be necessary. As common guidelines are still lacking on the evaluation of synthetic data in healthcare, further research may deal with a broader range of synthetic datasets, generated from other real-world datasets, to determine their specific similarity thresholds.

C. Related literature

In the past few years, some studies have explored different approaches for the generation and subsequent analysis of synthetic datasets in healthcare. Lu et al. [11] investigated the use of GANs to produce privacy-preserving synthetic data to circumvent possible privacy violation issues due to the release of publicly available datasets containing sensitive or identifying information. Specifically, correlation matrices were calculated to check whether the synthetic data preserved the original pairwise correlations between variables, and the similarity between the synthetic and original data distributions was assessed by evaluating the accuracy in a machine learning classification task, by considering the same conditions A, B, and C as described in Section V-A. In our study, we further expand the approach, by assessing whether the model trained on the original data is able to properly describe the synthetic data (condition D in the analysis of classification performance, subsection V-A). A recent study by El Emam et al [13] investigates the ability of a variety of utility metrics in evaluating 30 different health datasets and 3 different synthetic data generation methods including Bayesian networks, GANs, and sequential tree synthesis. According to the authors, the HD is the metric that best ranks the synthetic data generation methods based on prediction performance. Another interesting example of synthetic data validation is the study by Goncalves et al [12] that evaluates the quality of data generated from the cancer registry data from the Surveillance Epidemiology and End Results program of the US National Institutes of Health (NIH). Data were generated using Bayesian Networks and GANs and a set of different metrics were proposed, including utility metrics such as the Kullback-Leibler divergence, pairwise correlation difference, log-cluster metric, support coverage, as well as cross-classification (i.e., models trained on the original data only and tested on hold-out data from both original and generated data, and models trained on synthetic data only and tested on hold-out data from both original and generated data). However, even if a decision tree was used to compute the cross-classification metrics, the study did not address the rules extracted by the decision tree trained on the real and synthetic datasets. To our knowledge, no study so far has evaluated the quality of synthetic data by combining statistics, performance metrics and XAI-based measures. The results of this study confirm the potential value of XAI for assessing synthetic data qualitatively and quantitatively due to its ability to drive inspection of rules, thus clarifying the intrinsic mechanisms underlying the data.

VIII. Conclusion

This study demonstrates that XAI can provide additional insights in evaluating the quality of synthetic data, beyond the use of conventional utility metrics, in a hearing screening dataset. Specifically, a global similarity metric was introduced to assess the quality of synthetic data based on the similarity between the classification rule sets extracted from real and synthetic datasets. This metric allows for additional information about the synthetic dataset to be selected, when utility metrics do not allow for clear ranking. Moreover, XAI helps to highlight which input-output relationships are amplified in synthetic data and which ones may be neglected. Among the several XAI techniques available, the LLM was used in this study due to its ability to generate fully interpretable, rule-based models. However, future studies will be needed to investigate novel metrics based on other XAI approaches, for example post-hoc XAI techniques such as partial dependence plots or Shapley additive explanations. Further research is needed to investigate other datasets, including multivariate longitudinal data or time series from a large sample of subjects or biomedical signals to assess the generalizability of the proposed approach. Moreover, investigation of synthetic health data generated using other data generation algorithms (e.g., probabilistic models, classification-based imputation models, and different GAN algorithms) will be important to test whether XAI-derived metrics can be adapted to specific data generation algorithms and possibly used to assess the quality of synthetic data in real time, during the generation process. Providing real time feedback during the data generation process is one of the most promising goals to pursue as it could help improve the performance and efficiency of synthetic data generation methods.

Appendix I
Rules from synthetic and real WHISPER datasets

Table III shows the rules with covering higher than 15% obtained from the real dataset and from four out of 15 synthetic datasets , specifically the ones with better MMD (lower values, p-value > 0.05) and HD (lower values, < 0.40), as shown in Table I. The results of Table III are used to compute the coefficients shown in Table II (Section VI-A).

References

[1] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *npj Digit. Med.* 2^{nd} ed., vol. 3, 147, 2020. DOI: https://10.1038/s41746-020-00353-9

[2] J. de Benedetti, N. Oues, Z. Wang, P. Myles, and A. Tucker, " Practical Lessons from Generating Synthetic Healthcare Data with Bayesian Networks," in Koprinska I. et al. (eds). ECML PKDD 2020 Workshops. ECML PKDD 2020. *Communications in Computer and Information Science*, vol. 1323, Springer, Cham, 2020. DOI:$10.1007/978-3-030-65965-3_3$

[3] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data,"*Neurocomputing*, vol. 416, pp. 244–255, 2020. DOI: 10.1016/j.neucom.2019.12.136

[4] J. Yoon, L. N. Drumright and M. van der Schaar, "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378-2388, Aug. 2020. DOI: 10.1109/JBHI.2020.2980262

[5] B. M. Maweu, R. Shamsuddin, S. Dakshit and B. Prabhakaran, "Generating Healthcare Time Series Data for Improving Diagnostic Accuracy of Deep Neural Networks," in *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-15, 2021. DOI: 10.1109/TIM.2021.3077049

[6] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, M. Mongelli, "A Generative Adversarial Network (GAN) Technique for Internet of Medical Things Data," *Sensors*, vol. 21, no. 11:3726, 2021. DOI: 10.3390/s21113726

[7] N. Gulati, P. D. Kaur, "An argumentation enabled decision making approach for Fall Activity Recognition in Social IoT based Ambient Assisted Living systems," *Future Generation Computer Systems*, vol. 122, pp. 82-97, 2021. DOI: 10.1016/j.future.2021.04.005

[8] J. Zhai, J. Qi, and S. Zhang, "Imbalanced data classification based on diverse sample generation and classifier fusion,"*Int. J. Mach. Learn. & Cyber.*, 2021. DOI: 10.1007/s13042-021-01321-9

[9] Z. Wang, P. Myles, A. Tucker, "Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy," *Computational Intelligence*, vol. 37, pp. 819-851, 2021. DOI: 10.1111/coin.12427

[10] A. Borji, "Pros and cons of GAN evaluation measures,"*Computer Vision and Image Understanding*, vol. 179, pp. 41-65, 2019. DOI: 10.1016/j.cviu.2018.10.009

[11] P-H. Lu, P-C. Wang, and C-M. Yu, "Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network, in " *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics (WIMS2019)*, Association for Computing Machinery, Article 16, pp.1-6, 2019. DOI: 10.1145/3326467.3326474

[12] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A.P. Sales, "Generation and evaluation of synthetic patient data,"*BMC Med Res Methodol*, vol. 20, no. 108, 2020. DOI: 10.1186/s12874-020-00977-1

[13] K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna. "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study". *JMIR medical informatics*, vol. 10, no. 4, e35734, 2022. DOI:10.2196/35734

[14] L. Le Cam, and G.L. Yang. "Asymptotics in Statistics: Some Basic Concepts," New York, NY: Springer; 2000. DOI:10.1007/978-1-4612-1166-2

[15] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, " Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020. DOI: 10.1016/j.inffus.2019.12.012

[16] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, “Machine Learning Interpretability: A Survey on Methods and Metrics,” *Electronics*, vol. 8, no. 832, 2019. DOI: 10.3390/electronics8080832

[17] E. Khodabandehloo, D. Riboni, and A. Alimohammadi, “HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline,” *Future Generation Computer Systems*, vol. 116, pp. 168-189, 2021. DOI: 10.1016/j.future.2020.10.030

[18] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning-based prediction models in healthcare,” *WIREs Data Mining Knowl Discov.*, vol. 10, 2020. DOI: 10.1002/widm.1379.

[19] M. Lenatti, P.A. Moreno-Sanchez, E.M. Polo, M. Mollura, R.Barbieri, A. Paglialonga, “Evaluation of machine learning algorithms to detect hearing loss from a speech-in-noise screening test,” *The American Journal of Audiology* vol 31(3S), pp. 961-979, 2022. DOI: 10.1044/2022_aja-21-00194.

[20] A. Paglialonga, E.M. Polo, M. Zanet, G. Rocco, T. van Waterschoot, and R. Barbieri, “An automated speech-in-noise test for remote testing: development and preliminary evaluation,” *Am J Audiol.*, vol. 29, pp. 564-576, 2020. DOI:$10.1044/2020\_aja - 19 - 00071$

[21] World Health Organization “World report on hearing,”*Geneva: World Health Organization*, 2021. Licence: CC BY-NC-SA 3.0 IGO. Available: https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/highlighting-priorities-for-ear-and-hearing-care

[22] A. Davis and P. Smith, “Adult hearing screening: Health policy issues—What happens next?,” *American Journal of Audiology*, vol. 22, no. 1, pp. 167-170, 2013. DOI: 10.1044/1059-0889(2013/12-0062)

[23] M.C.J. Leensen, J. A. P. M. de Laat, W. A. Dreschler, “Speech-in-noise screening tests by internet, Part 1: Test evaluation for noise-induced hearing loss identification,” *International Journal of Audiology*, vol. 50, no. 11, pp. 823-834, 2011. DOI: 10.3109/14992027.2011.595016

[24] World Health Organization (WHO), “Report of the informal working group on prevention of deafness and hearing impairment programme planning: Geneva,” June, 18-21, 1991. Available: https://apps.who.int/iris/handle/10665/58839

[25] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS2014)*, vol. 2, pp. 2672-2680, 2014.

[26] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, “On Data Augmentation for GAN Training,” *IEEE Trans. on Image Process*, vol. 20, pp. 1882-1897, 2021. DOI: 10.1109/tip.2021.3049346

[27] M.S. Munia, M. Nourani, and S. Houari, “Biosignal oversampling using Wasserstein Generative Adversarial Network,” in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, 2020. DOI: 10.1109/ichi48887.2020.9374315

[28] D. Hazra, and Y.-C. Byun, “SynSigGAN: Generative Adversarial Networks for synthetic biomedical signal generation,” *Biology*, vol. 9, no. 441, 2020. DOI:10.3390/biology9120441

[29] R. Atienza, “Advanced Deep Learning with Keras: Apply Deep Learning Techniques, Autoencoders, GANs, Variational Autoencoders, Deep Reinforcement Learning, Policy Gradients, and More,” Packt Publishing, 2018.

[30] Y. Zhou, B. Wang, X. He, S. Cui, and L. Shao, “DR-GAN: Conditional Generative Adversarial Network for Fine-Grained Lesion Synthesis on Diabetic Retinopathy Images, in ” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 56-66, Jan. 2022. DOI: 10.1109/JBHI.2020.3045475

[31] A. Gretton, K.M. Borgwardt, M. Rasch, B. Schoelkopf, and A.J. Smola, “Kernel method for the two-sample-problem, in ” *Advances in Neural Information Processing Systems (NIPS’06)*, MIT Press, pp. 513-520, 2006.

[32] I. Kim, A. Ramdas, A. Singh, and L. Wasserman, “Classification accuracy as a proxy for two-sample testing,” *Ann. Statist.*, vol. 49, pp. 411-434, 2021. DOI: 10.1214/20-aos1962

[33] H. Cai, B. Goggin, and Q. Jiang, “Two-sample test based on classification probability,” *Stat Anal Data Min: The ASA Data Sci Journal*, vol. 13, pp. 5-13, 2019. DOI:10.1002/sam.11438

[34] M.Muselli, “Switching Neural Networks: A New Connectionist Model for Classification, in: Neural Nets,” in *Neural Nets*,Springer Berlin Heidelberg, pp. 23-30, 2006. DOI: $10.1007/11731177_4$.

[35] M. Muselli, E. Ferrari, “Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, pp. 37-50, 2011. DOI: 10.1109/tkde.2009.206

[36] Y. Goldberg, *Neural Network Methods for Natural Language Processing*, vol. 37, Morgan & Claypool , San Rafael, CA, 2017. ISBN: 978-1-62705-298-6

[37] K. Juluru, H.H. Shih, K.N. Keshava Murthy, and P. Elnajjar, “Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists,” *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 41, no. 5, pp. 1420-1426, 2021. DOI: 10.1148/rg.2021210025

[38] D. Polap, M. Wlodarczyk-Sielicka, “Classification of Non-Conventional Ships Using a Neural Bag-Of-Words Mechanism,” *Sensors (Basel)*, vol. 20, no. 6, 2020. DOI: 10.3390/s20061608

[39] S. Narteni, M. Ferretti, V. Rampa, M. Mongelli, M. “Bag-of-Words Similarity in eXplainable AI”. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 543, 2022. Springer, Cham. DOI: 10.1007/978-3-031-16078-3_58

[40] S. Hirano and S. Tsumoto, “Detection of Differences between Syntactic and Semantic Similarities,”Tsumoto S., Słowiński R., Komorowski J., Grzymała-Busse J.W. (eds), *Rough Sets and Current Trends in Computing, RSCTC 2004, Lecture Notes in Computer Science*, vol. 3066. Springer, Berlin, Heidelberg, 2004.

[41] J.L Wu, X. Xiao, L.C. Yu, S.Z. Ye, K.R. Lai, “Using an analogical reasoning framework to infer language patterns for negative life events,” *BMC Med Inform Decis Mak*, vol. 19, no. 1:173, 2019. DOI: 10.1186/s12911-019-0895-8

[42] L.E. Humes, “Understanding the Speech-Understanding Problems of Older Adults,” *Am J Audiol.* , vol. 22, pp. 303-305, 2013. DOI: 10.1044/1059-0889(2013/12-0066)

[43] M. Zanet, E.M. Polo, M.Lenatti, T. van Waterschoot, M. Mongelli, R. Barbieri, and A. Paglialonga, “Evaluation of a Novel Speech-in-Noise Test for Hearing Screening: Classification Performance and Transducers’ Characteristics,” in *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4300-4307, Dec. 2021. DOI: 10.1109/JBHI.2021.3100368

[44] R.V. Zicari, J. Brusseau, S.N. Blomberg, H.C. Christensen, M. Coffee, M.B. Ganapini et al. “On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls ,” in *Frontiers in Human Dynamics*, vol.3, 2021. DOI: 10.3389/fhumd.2021.673104

TABLE III
Rules with higher covering extracted by the LLM from WHISPER dataset and from the synthetic datasets #8, #9, #13, and #15, divided by output classes: 'no HL' and 'HL'

Data	#	Rule conditions	C(%)
real	$R_{r,noHL}1$	$Age \leq 52$	60.8
	$R_{r,noHL}2$	$(-18.82 < SRT \leq -7.35) \wedge (\#correct > 96)$	46.8
	$R_{r,noHL}3$	$SRT \leq -16.11$	21.74
	$R_{r,HL}1$	$\#correct \leq 59$	52.17
	$R_{r,HL}2$	$(-10.30 < SRT \leq -0.24) \wedge (Age > 66)$	36.96
	$R_{r,HL}3$	$(-15.46 < SRT \leq 5.78) \wedge (47 < Age \leq 74) \wedge (\#correct \leq 68)$	19.56
	$R_{r,HL}4$	$-16.38 < SRT \leq -11.80$	17.39
#8	$R_{8,noHL}1$	$Age \leq 48$	52.82
	$R_{8,noHL}2$	$(-12.53 < SRT \leq 24.95) \wedge (25 < Age \leq 86) \wedge (58 < \#correct \leq 113)$	50.28
	$R_{8,HL}1$	$(SRT > -9.46) \wedge (48 < Age \leq 85) \wedge (\#correct < 108)$	78.47
	$R_{8,HL}2$	$(Age > 49) \wedge (\#correct \leq 63)$	51.79
#9	$R_{9,noHL}1$	$Age \leq 66$	73.03
	$R_{9,noHL}2$	$Age \leq 47$	53.92
	$R_{9,noHL}3$	$(SRT \leq -6.21) \wedge (48 < Age \leq 73.44) \wedge (\#correct > 62)$	39.46
	$R_{9,HL}1$	$(Age > 47) \wedge (\#correct \leq 62.61)$	80.36
	$R_{9,HL}2$	$Age > 65$	63.26
#13	$R_{13,noHL}1$	$(SRT \leq -13.80) \wedge (60 < \#correct \leq 103.32)$	52.03
	$R_{13,noHL}2$	$Age \leq 47$	50.76
	$R_{13,noHL}3$	$(-15.7 < SRT \leq -4.34) \wedge (42 < Age \leq 85) \wedge (72 < \#correct \leq 103)$	18.53
	$R_{13,noHL}4$	$(-14.03 < SRT \leq -5.34) \wedge (42 < Age \leq 65) \wedge (\#correct > 53)$	18.27
	$R_{13,HL}1$	$(SRT > -7.04) \wedge (Age > 43)$	67.24
	$R_{13,HL}2$	$(Age > 67) \wedge (\#correct \leq 79)$	65.76
	$R_{13,HL}3$	$Age > 78$	30.29
#15	$R_{15,noHL}1$	$(SRT \leq -4.3) \wedge (\#correct > 72.43)$	54.72
	$R_{15,noHL}2$	$Age \leq 47$	48.01
	$R_{15,noHL}3$	$(-17.05 < SRT \leq -9.49) \wedge (Age \leq 64) \wedge (\#correct > 55)$	36.07
	$R_{15,noHL}4$	$(-18.55 < SRT \leq -16.48) \wedge (Age > 45) \wedge (\#correct > 63)$	32.83
	$R_{15,noHL}5$	$(SRT \leq -17.75)$	26.37
	$R_{15,HL}1$	$(SRT > -9.49) \wedge (47.35 < Age \leq 86.13)$	71.36
	$R_{15,HL}2$	$(-18.77 < SRT \leq -2.31) \wedge (49 < Age \leq 86) \wedge (\#correct \leq 65)$	37.18