Research Article

Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss From a Speech-in-Noise Screening Test

Marta Lenatti,[a] Pedro A. Moreno-Sánchez,[b,c] Edoardo M. Polo,[d] Maximiliano Mollura,[e] Riccardo Barbieri,[e] and Alessia Paglialonga[a]

[a] Institute of Electronics, Information Engineering and Telecommunications, National Research Council of Italy, Milan [b] School of Health Care and Social Work, Seinäjoki University of Applied Sciences, Finland [c] Faculty of Medicine and Health Technology, Tampere University, Seinäjoki, Finland [d] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Italy [e] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Article History:
Received September 14, 2021
Revision received December 24, 2021
Accepted May 8, 2022

Editor-in-Chief: Ryan W. McCreery
Editor: Valeriy Shafiro

https://doi.org/10.1044/2022_AJA-21-00194

ABSTRACT

Purpose: The aim of this study was to analyze the performance of multivariate machine learning (ML) models applied to a speech-in-noise hearing screening test and investigate the contribution of the measured features toward hearing loss detection using explainability techniques.
Method: Seven different ML techniques, including transparent (i.e., decision tree and logistic regression) and opaque (e.g., random forest) models, were trained and evaluated on a data set including 215 tested ears (99 with hearing loss of mild degree or higher and 116 with no hearing loss). Post hoc explainability techniques were applied to highlight the role of each feature in predicting hearing loss.
Results: Random forest (accuracy = .85, sensitivity = .86, specificity = .85, precision = .84) performed, on average, better than decision tree (accuracy = .82, sensitivity = .84, specificity = .80, precision = .79). Support vector machine, logistic regression, and gradient boosting had similar performance as random forest. According to post hoc explainability analysis on models generated using random forest, the features with the highest relevance in predicting hearing loss were age, number and percentage of correct responses, and average reaction time, whereas the total test time had the lowest relevance.
Conclusions: This study demonstrates that a multivariate approach can help detect hearing loss with satisfactory performance. Further research on a bigger sample and using more complex ML algorithms and explainability techniques is needed to fully investigate the role of input features (including additional features such as risk factors and individual responses to low-/high-frequency stimuli) in predicting hearing loss.

Age-related hearing loss has a high prevalence, and if left untreated, it can lead to cognitive decline, social isolation, and mental health problems that can translate from a health care system perspective into an increased access to care and higher costs (Dalton et al., 2003; Davies et al., 2017; Reed et al., 2019). Despite this, hearing loss is frequently considered an inevitable component of aging. Adult hearing screening can help increase consciousness and subsequent early detection of hearing loss and help prevent the effects of undiagnosed hearing loss on quality of life (Davis et al., 2007; Feltner et al., 2021).

An important recent trend in hearing health care is related to the development of methods for adult hearing screening, particularly those based on speech-in-noise testing, delivered either locally or at a distance (Blamey et al., 2015; Leensen et al., 2011; Paglialonga et al., 2014, 2020;

Correspondence to Marta Lenatti: marta.lenatti@ieiit.cnr.it. **Publisher Note:** This article is part of the Special Issue: 5th International Meeting on Internet and Audiology. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Smits et al., 2004; Watson et al., 2012). Speech-in-noise testing has become a popular approach for adult hearing screening as the first difficulties experienced by adults with hearing loss are typically related to understanding speech in noisy environments. Furthermore, speech-in-noise tests can be performed in nonclinical settings without a stringent need for calibrated equipment or transducers (e.g., Smits et al., 2004). Typically, the outcome of a speech-in-noise screening test is the speech recognition threshold (SRT; Leensen et al., 2011; Paglialonga et al., 2020; Smits et al., 2004; Watson et al., 2012) or the number/percentage of correct responses (Blamey et al., 2015; Paglialonga et al., 2013). Therefore, the screening outcome is commonly determined based on a single variable. Specifically, hearing loss is usually detected whenever the SRT or the number/percentage of correct responses is above or below a certain cutoff value, respectively. The use of these speech-in-noise measures to predict hearing loss, as assessed using pure-tone thresholds, is frequently reported in the literature, particularly in the area of hearing screening tests (e.g., Blamey et al., 2015; Paglialonga et al., 2014; Smits et al., 2004). Although widespread, such a univariate approach can present limitations. First, there is a well-known mismatch between pure-tone thresholds and SRTs as individuals with normal pure-tone thresholds may have difficulties in speech understanding and, vice versa, individuals with hearing loss may be able to reach satisfactory speech recognition performance (Humes, 2013; Killion & Niquette, 2000). Second, other features, in addition to SRT, might be valid predictors of hearing loss, for example, the subject's age or the average reaction time (Humes, 2013; Nuesse et al., 2018; Polo, Zanet, Lenatti, et al., 2021; Polo, Zanet, Paglialonga, & Barbieri, 2021). Nevertheless, possible multivariate classification methods for identifying hearing loss from speech-in-noise tests have not been systematically explored yet.

In recent years, machine learning (ML) has gained increasing popularity in a wide variety of domains, not least in the medical field. As ML models can uncover hidden patterns that are not immediately visible to the clinician's eye, they can support clinical decision making and may help in the process of automating screening tests, potentially allowing for their widespread adoption (Kumar, 2019). In general, ML is envisioned as an appropriate solution to discover latent correlations and can be used to assist medical personnel to identify existing conditions or patients at risk of developing disease. ML offers advantages not only for predictive and prescriptive analytics but also for feature ranking and model interpretation, that is, for a deeper understanding of the role of input features in determining the model output, and as such, it can help clinicians in interpreting the model outcomes. There is growing interest in the area of explainable artificial intelligence (XAI) as, broadly speaking, model explanations (in a variety of formats) can support health care experts in making data-driven decisions and therefore provide more personalized and trustworthy treatments (Belle & Papantonis, 2021; Vaccari et al., 2021; Vilone & Longo, 2021). Explainability approaches can be broadly grouped into "transparent" models (e.g., decision trees [DTs] and logistic regression [LR]) and "opaque" models (e.g., random forest [RF] and support vector machine [SVM]) combined with post hoc explainability techniques (Belle & Papantonis, 2021). The former provide insights into the procedure they carry out to generate predictions, for example, in terms of hierarchical rules (e.g., DT) or numerical coefficients, which combine with the input features to provide the outputs (e.g., LR). Conversely, the latter rely on more complex mechanisms like high-dimensional representations and data transformations (e.g., SVM) or combinations of models (e.g., RF), and therefore, they are not inherently explainable (further details are reported in the ML Approach section). In this context, post hoc explainability techniques applied to opaque models can help define a suitable balance in terms of explainability and accuracy (Moreno-Sanchez, 2020). A variety of post hoc explainability techniques have been developed to increase transparency and therefore understanding of the logic underlying the models as, based on the context and application, different formats of explanations can be considered, for example, numerical, rule-based, textual, visual, or a mixture of formats (Vilone & Longo, 2021). Examples include implicit feature importance scores, SHapley Additive exPlanations (SHAP), and partial dependence plots (PDPs; Carvalho et al., 2019; Friedman, 2001).

Recently, we have developed and validated a new tool for speech-in-noise testing that is able to extract a number of features related to the individual performance, for example, SRT, average reaction time, age, total test time, and number and percentage of correct responses (Paglialonga et al., 2020). Preliminary results showed that by introducing ML models, the tool is able to identify the presence of hearing loss with a satisfactory level of accuracy. Indeed, preliminary results showed that multivariate classifiers based on, for example, LR, SVM, *k*-nearest neighbors (KNN), or RF can be more accurate than a univariate classifier based on the SRT to identify hearing loss of slight/mild degree or higher (Zanet et al., 2021), as defined by the former World Health Organization (WHO) definition (WHO, 1991).

The aim of this study was to (a) investigate different ML explainability approaches, including transparent and opaque models, in combination with post hoc explainability techniques, to identify hearing loss in adults using the newer WHO definition, introduced on March 2021 (WHO, 2021a, 2021b), and (b) characterize the features extracted from the speech-in-noise screening test in terms of their contribution to the prediction of hearing loss.

Such an XAI-based approach could help generate more knowledge about the features that can effectively characterize hearing sensitivity in adults and, in turn, help building accurate methods for hearing loss identification, thus being of substantial interest to researchers and developers of audiological instrumentation.

Method

Participants

Data were collected from a cohort of 207 unscreened adults (66 men and 141 women; $M_{\text{age}} = 52$ years, $SD = 20$, range: 20–89 years) of varying native languages (Italian: 170 subjects; English: 12 subjects; Arabic: six subjects; Spanish: six subjects; French, German, Somali, Albanian, Filipino, Moroccan Arabic, Igbo, and Efik: less than four subjects) during hearing loss prevention and awareness events for the public held at not-for-profit organizations and lifelong learning institutes. Inclusion criteria were age of ≥ 18 years and ability to use a mouse and interact with the platform. Exclusion criteria (ear-specific) were hearing aids or implantable devices. Participants were given the option to choose in which ear(s) to perform the test. In the experiment, only eight subjects were tested in both ears (i.e., yielding a data set of 215 records).

Procedure

Pure-tone testing was performed at 0.5, 1, 2, and 4 kHz using a clinical audiometer (Amplaid 177+, Amplifon with TDH49 headphones). Speech-in-noise testing was performed by using a recently validated user-operated screening test (Paglialonga et al., 2020; Polo, Zanet, Lenatti, et al., 2021; Zanet et al., 2021). The test is based on a three-alternative multiple-choice speech recognition task, that is, with three written alternatives displayed on the screen, and delivers speech stimuli at varying signal-to-noise ratio (SNR) based on a newly developed one-up/three-down staircase algorithm (Paglialonga et al., 2020; Zanet et al., 2019). The test stimuli are 12 meaningless vowel–consonant–vowel (VCV) syllables in the context of the vowel /a/ (e.g., ata and asa) recorded from a professional native English speaker (Paglialonga et al., 2014, 2020). The rationale for using meaningless stimuli in a multiple-choice task was to limit the influence of education and literacy in a way that only basic reading skills and minimal knowledge of the Latin alphabet are required to perform the test (Cooke et al., 2010; Mattys et al., 2009; Paglialonga et al., 2020). The spoken stimuli from a native English speaker were selected from a multilingual corpus from an earlier VCV-based test (e.g., Paglialonga et al., 2014; Vaez et al., 2014). Specifically, the English set of VCVs was selected based on the analysis of speech recognition performance estimated from a combination of listening tests and nonintrusive intelligibility measures in a way to reduce possible effects on speech recognition in nonnative listeners (Rocco, 2018). To further limit the possible influence of native language on individual performance, the 12 consonants used in the test were selected among those that have similar pronunciation and transcription across some of the most widely spoken languages worldwide such as English, Spanish, French, Portuguese, German, and Italian ("b" and "v" were not used as they may have the same pronunciation, e.g., in Spanish). A detailed description of the speech-in-noise testing procedure can be found in the studies of Paglialonga et al. (2020) and Zanet et al. (2021).

In the initial phase of the study, the first 148 participants (no exclusion criteria were applied) also completed the Hearing Handicap Inventory for the Elderly–Screening Version (HHIE-S; Ventry & Weinstein, 1983) to gain a deeper insight into the relationship between hearing loss, as predicted using the proposed screening system, and the perceived hearing handicap.

Both pure-tone and speech-in-noise testing were performed in low environmental noise settings in dedicated rooms at the sites hosting the hearing loss prevention and awareness events. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion No. 2/2019, February 19, 2019). Participants were informed about the protocol and took part in this experiment on a voluntary basis.

Data Analysis

The data set used in this study includes six input features extracted from speech-in-noise testing and one output feature, that is, the presence or absence of hearing loss, as determined by the pure-tone average (PTA), that is, the average value of pure-tone thresholds measured at 0.5, 1, 2, and 4 kHz. Differently from our previous studies (Paglialonga et al., 2020; Zanet et al., 2021), in this study, the ears tested were classified using the updated WHO definition of slight/mild hearing impairment introduced on March 2021 (WHO, 2021a, 2021b). Specifically, each ear (i.e., each record in the data set) was classified as having hearing loss (class "HL") when the PTA was higher than 20 dB HL, whereas it was classified as not having hearing loss (class "no HL") if the PTA was lower than or equal to 20 dB HL. Based on this definition, 116 out of 215 records (54%) were classified in the no-HL class (PTA: M = 8.74 dB HL, SD = 8.34), whereas the remaining 99 ears (46%) were classified in the HL class (PTA: M = 35.32 dB HL, SD = 10.34), resulting in a quite balanced data set.

The six input features extracted upon completion of the test in each ear comprise SRT (Mdn = −12.11 ± 8.3 dB, range: −20.4 to +19.2 dB), number of correct responses (*Mdn* = 68 ± 14.9 correct responses, range: 28–113), percentage of correct responses (*Mdn* = 90% ± 4.3%, range: 55.4%–93.3%), average reaction time (i.e., time needed to provide a response, averaged across the staircase procedure; *Mdn* = 1.8 ± 0.78 s, range: 0.91–6.11 s), total test time (*Mdn* = 239 ± 59.8 s, range: 145–497 s), and subject's age (*Mdn* = 59 ± 20.9 years, range: 20–89 years). It should be noted that the number of correct responses codifies a slightly different information than the percentage of correct responses (i.e., the ratio number of correct/number of trials) since the number of trials in the adaptive speech-in-noise test is not fixed.

ML Approach

For the purpose of building ML models for classification of ears into HL or no HL, the data set was randomly split into a training set including 80% of the sample (172 records) and a test set (unseen data) including the remaining 20% (43 ears). Stratification was applied to maintain the same percentage of records in the two classes of the original data set in the training and test partitions. Before applying ML algorithms, data were standardized (based on the training data), that is, transformed to have zero mean and unit variance to limit the influence of features defined on different value ranges on model training (e.g., Luor, 2015).

Due to the relatively small size of the data set (215 records, six input features), fivefold cross-validation was introduced on the training set to partially reduce the influence of the selected partition on the trained model. For the same reason, the performance of ML algorithms was evaluated independently on 50 different iterations of model training and testing on 50 randomly generated realizations of the training and test sets to address the variability of the classification model due to changes in the underlying data.

Seven ML algorithms have been investigated, namely, four of the most widely used approaches (DT, LR, SVM, and KNN; James et al., 2013) and three ensemble methods (ensemble LR [ELR], RF, and gradient boosting [GB]; Sagi & Rokach, 2018). DTs are transparent models that can be described by a set of m intelligible rules r_k, ($k = 1, \ldots, m$), in the format *if (premise) then (consequence)*, where *premise* is a logical product of n conditions c_{ik}, with $ik = 1, \ldots, nk$, and *consequence* provides a class assignment for the output (e.g., class HL). A DT builds a tree-structured classifier, where the internal nodes represent the features of a data set, the branches represent the splitting rules, and each leaf node represents the outcome of the model. Following preliminary investigations, we built DT models with a maximum achievable depth equal to four levels, using the Gini index to quantify the purity of classification in a node. LR operates similarly to linear regression but is used for classification of binary targets. It builds models by mapping the predicted values in probability values ranging from 0 to 1 by means of a sigmoid (i.e., S-shaped) function under the assumption that the output is a linear combination of the predictor variables. Transparency of LR is related to the fact that the model coefficients can be interpreted as weighting factors on input features and therefore can be used to understand the model's prediction mechanism. SVM is an algorithm that aims to find a hyperplane, that is, a multivariate decision boundary in the n-dimensional features space, able to discriminate the observations in the best way by means of kernel functions. In this study, SVMs with linear kernels were addressed. KNN is a distance-based ML algorithm built on the idea that multivariate observations that are close to each other in the n-dimensional features space belong to the same class. The main parameter of a KNN is the number of neighbors k (set here to the commonly used value $k = 5$) used to classify data points. Ensemble techniques are ML techniques that combine decisions from multiple models, called base models or weak learners, to obtain better classification performance with respect to a single model, trying to reduce its bias and/or variance (Dietterich, 2000). Particularly, RF and GB are homogeneous (i.e., based on a single kind of weak learner), tree-based techniques, whereas the ELR is a heterogeneous technique (i.e., based on different kinds of ML models as weak learners). RF is composed of several DTs trained in parallel on different subsets of the original data set, drawn by means of bootstrapping. The final prediction of the RF model is obtained from the average of DT predictions. GB is an additive sequential approach based on boosting where a new weak learner is added at each iteration, trying to correct and improve the previous model. Following preliminary analyses, the number of estimators (i.e., DT) for both RF and GB was set to 50 since this number consistently yielded better training accuracy. ELR is based on two-level processing. The first level includes fitting DT, SVM, LR, and KNN as base models. The second level includes an LR trained on the predicted probabilities generated by the base models and used to predict new data. Post hoc explainability techniques are applied to the algorithm exhibiting the highest classification performance, as determined by the analysis of the following measures, averaged across 50 iterations: accuracy on the training set, accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and precision on the test set. All the ML algorithms were implemented using the scikit-learn Python library (Python Version 3.7.10).

The Lilliefors test was performed to check for normality of the distributions of the performance metrics across the 50 iterations. The nonparametric Kruskal–Wallis test was performed to assess possible differences in performance metrics among the different methods. When significant differences in the median values of classification metrics were observed, post hoc nonparametric multiple-comparison tests with Bonferroni correction were performed. A significance level of α = .05 was considered.

The distributions of HHIE-S scores measured in the predicted output classes (HL vs. no HL), as identified using the algorithm exhibiting the highest classification performance, are compared. Possible differences in median values between the two classes are assessed using the nonparametric Kruskal–Wallis test.

Post Hoc Explainability Techniques

Different post hoc explainability techniques have been used to assess the importance of each feature in the model output for both individual predictions and overall binary classification of the sample. First, implicit feature importance is considered. Different ML algorithms provide an implicit measure of feature importance that is determined by the feature's coefficients assigned in the mathematical formulation of the model. This implicit feature importance is inherently model dependent, and it provides different information for different models. For example, in RF, the implicit importance of a given feature denotes the mean decrease of impurity, as measured by the Gini impurity index, observed in the base classifier's nodes of the different DT estimators. Specifically, a node is defined as "pure" whenever it contains only instances of a certain class (Gini index = 0), whereas a node is defined as "impure" if it contains instances equally distributed across different output classes (Gini index = 0.5). Thus, the bigger the observed decrease of impurity for a certain feature, the more important is that feature in determining the output of the model.

In addition to implicit feature importance, model-agnostic post hoc explainability techniques can be addressed. These techniques can be applied to any ML model to discover the importance of features in the model's output and to explore the influence of the values of a given feature in the final probability of the classification. Specifically, the following model-agnostic post hoc techniques were used: feature permutation importance, SHAP, and PDP.

The feature permutation importance technique (Fisher et al., 2019) is based on measuring the increase in the prediction error of a model, that is, the decrease in classification accuracy, when the values of a specific feature are shuffled, therefore breaking the relationship between that feature and the true outcome. Therefore, the larger the decrease in accuracy observed by shuffling the values of a certain feature, the more important the feature is because the model relies on the feature's values for estimating the prediction accurately. Vice versa, a feature is denoted as unimportant if the accuracy of the model is minimally altered when the values of the feature are shuffled.

The SHAP technique provides explainability for individual predictions by computing an additive (positive or negative) measure of feature importance to the predicted outcome by applying coalitional game theory (Lundberg et al., 2018). Thus, SHAP quantifies the contribution that each feature brings to the prediction made by the model starting from the initial proportion of classification, for example, the initial prevalence of a given class. In this study, we used the SHAP technique in the form of the "waterfall" visualization, where both the computed additive contributions of each feature and their direction (positive or negative) toward the output predicted class are shown.

PDPs show the marginal effect of a given feature on the predicted outcome over the range of its observed values. The PDP works by making predictions for each instance of the data set across a range of values of a specific feature while all other features are kept constant (Friedman, 2001). As a result, PDPs visualize the overall probability of a given model output for every value of a certain feature, all the other features being equal. As such, PDPs are dependent on the initial proportion of output classes that represent a baseline for interpreting probability values. For example, values of PDP probability below the initial proportion of a given class for a certain range of feature values suggest that the probability of classifying an instance in that class is decreasing in that range of values. Therefore, implicit feature importance and feature permutation importance measure overall importance irrespectively of the output class, whereas SHAP and PDPs provide information about the direction of the influence of a given feature with respect to the two output classes and additional information about the values that contribute more prominently to classification into a specific class.

Results

Classification Performance

As a benchmark for multivariate classification, univariate classification performance obtained using an LR model trained on the whole data set is shown in Table 1. The table shows the overall classification performance of each of the six input features in determining the output class, as measured by the AUC. Based on its definition, a model that misclassifies all records yields AUC = 0, a model that correctly classifies all records yields AUC = 1, whereas AUC = .5 represents the chance level. The cutoffs represent the best discriminating thresholds according to the receiver operating characteristic curves. Among the six input features addressed here, age is the one that better discriminates between the two classes (AUC > .90, accuracy = .85). SRT, the number of correct responses, the percentage of correct responses, and average reaction time show lower but still good discrimination capabilities

Table 1. Classification performance of each single feature on the whole data set in terms of area under the receiver operating curve (AUC), accuracy, and cutoff value (i.e., the best discrimination threshold for splitting the two output classes), computed using a logistic regression model trained on the whole data set.

Variable	SRT	Age	#correct	%correct	Average reaction time	Total test time	All features
AUC	.79	.92	.74	.78	.79	.58	.94
Accuracy	.73	.85	.69	.71	.77	.34	.88
Cutoff	−7.48 dB SNR	53 years	63	90.28%	1.83 s	253 s	

Note. SRT = speech recognition threshold; #correct = number of correct responses; %correct = percentage of correct responses.

(AUC > .70, accuracy > .69). Conversely, total test time shows poor discrimination abilities, with AUC close to the chance level (AUC = .58) and very low accuracy (.34). The multivariate model obtained using LR on the full set of six features has improved performance compared to the univariate ones, specifically exhibiting AUC = .94 and accuracy = .88.

Table 2 shows the average classification performance over 50 iterations of the seven ML methods addressed in the study. Overall, the average performance of the different models is similar, with both training and test accuracies higher than .82 and no remarkable differences in performance between the training set and the test set (i.e., no sign of overfitting). Specificity and sensitivity are greater than .80 and have similar values, indicating good performance in correctly discriminating both the HL and no-HL classes. The average AUC is above .90, except for the DT (AUC = .85), indicating, overall, very good classification performance, not far from the ideal performance (i.e., AUC = 1). The standard deviation values in Table 2 indicate the variability of the performance measures when considering different training and test sets in a relatively small data set. As it can be noticed, the observed standard deviations are relatively low for each of the computed metrics. Specifically, the standard deviation is smaller than 0.09 for all the metrics except for sensitivity for which it reaches values up to 0.12 (when considering ELR and GB). The Lilliefors test for normality revealed that all the distributions of metrics over 50 iterations were not normal; therefore, the nonparametric Kruskal–Wallis test was performed. No significant differences in sensitivity are observed (p = .30), and the average values are in the range 84%–87%, indicating, overall, a remarkable ability to correctly classify ears with hearing loss. However, the different algorithms show significantly different performances in terms of accuracy on the test set (p = .01), AUC (p < .001), specificity (p = .001), and precision (p < .001). Post hoc pairwise tests indicate that DT has significantly lower test accuracy (p = .01), specificity (p = .02), and precision (p = .003) than RF and significant lower AUC than all the other methods (SVM, LR, ELR, RF, GB: p < .001; KNN: p = .001). Also, KNN has a significantly lower AUC than LR (p = .01) and RF (p = .004) and significantly lower specificity (p = .01) and precision (p = .01) than RF.

Overall, DT and KNN exhibit lower performance in detecting hearing loss than the other ML models considered, whereas no statistical differences were found between SVM, LR, GB, and RF in terms of classification metrics. The algorithm that shows the highest performance measures and the best trade-off between sensitivity (.86) and specificity (.85) is RF.

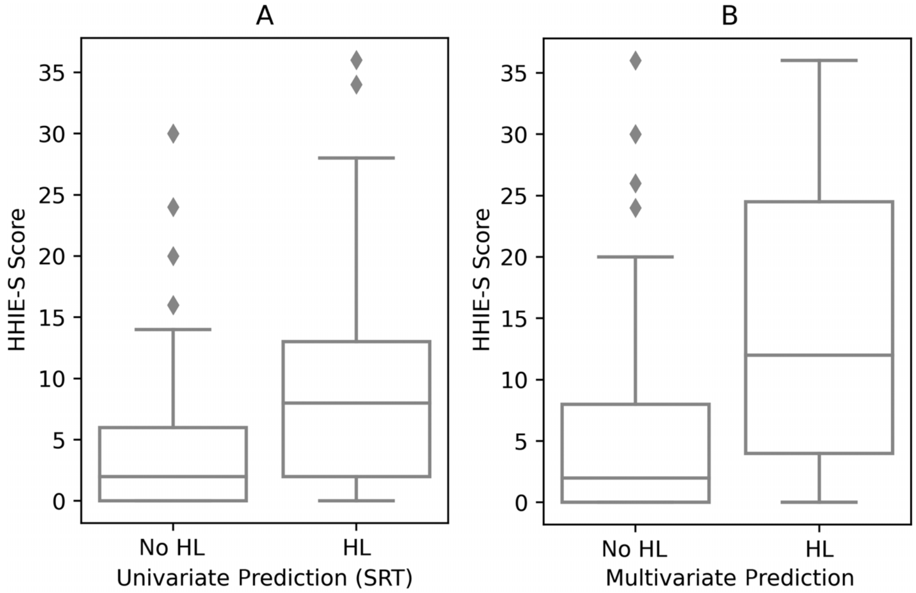
Figure 1 reports the distributions of the HHIE-S scores measured in the HL and no-HL classes, as determined

Table 2. Classification performance measures in terms of mean and standard deviation of different machine-learning techniques over 50 iterations of training and test set.

	Training accuracy		Test accuracy		Sensitivity		Specificity		AUC		Precision	
Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
DT	0.82	0.03	0.82	0.05	0.84	0.08	0.80	0.07	0.85	0.04	0.79	0.04
KNN	0.83	0.02	0.83	0.04	0.87	0.09	0.80	0.08	0.90	0.04	0.80	0.06
SVM	0.85	0.02	0.84	0.05	0.84	0.1	0.84	0.07	0.93	0.03	0.83	0.06
LR	0.85	0.01	0.84	0.05	0.85	0.08	0.84	0.08	0.93	0.03	0.83	0.07
ELR	0.91	0.02	0.84	0.05	0.86	0.12	0.83	0.07	0.92	0.04	0.82	0.06
RF	0.85	0.02	0.85	0.04	0.86	0.09	0.85	0.08	0.93	0.03	0.84	0.07
GB	0.85	0.02	0.84	0.05	0.86	0.12	0.83	0.07	0.92	0.04	0.82	0.06

Note. AUC = area under the receiver operating characteristic curve; DT = decision tree; KNN = *k*-nearest neighbors; SVM = support vector machine; LR = logistic regression; ELR = ensemble logistic regression; RF = random forest; GB = gradient boosting.

Figure 1. Analysis of the Hearing Handicap Inventory for the Elderly–Screening Version (HHIE-S) score with respect to predicted output class (HL vs. no HL) according to (A) a univariate classifier based on speech recognition threshold (SRT) and (B) a multivariate classifier based on the full set of features (*N* = 148 participants). HL = hearing loss.

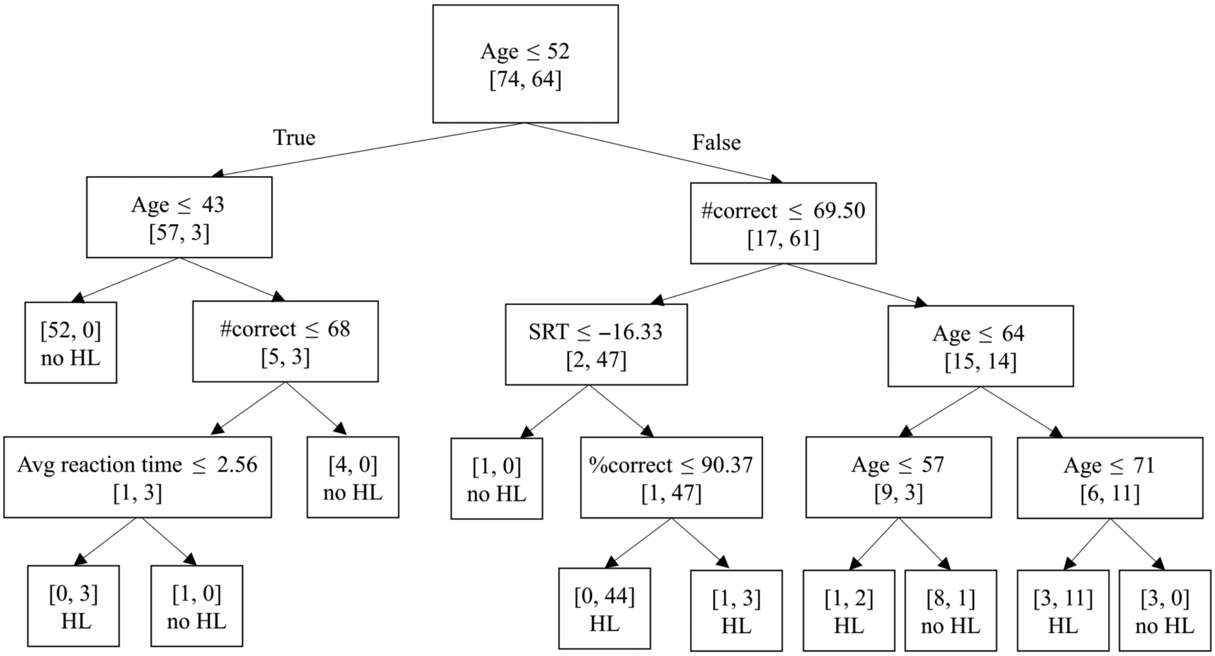


by (a) the univariate classifier based on SRT and (b) the multivariate RF model with the highest classification performance, respectively. The nonparametric Kruskal–Wallis test indicates that the median HHIE-S scores measured in the predicted no-HL class are significantly lower than those measured in the predicted HL class, considering both the SRT cutoff (predicted no HL: *Mdn* = 2, min = 0, max = 30; predicted HL: *Mdn* = 8, min = 0, max = 36; $p < .001$) and the multivariate classifier (predicted no HL: *Mdn* = 2, min = 0, max = 36; predicted HL: *Mdn* = 12, min = 0, max = 36; $p < .001$). Considering the criterion used to identify hearing handicap based on the HHIE-S, 40 out of 148 participants had self-reported hearing handicap (HHIE-S score ≥ 10). The hearing handicap class was correctly predicted in 107 out of 148 participants using the univariate classifier based on SRT (21 with hearing handicap and 86 without) and in 110 out of 148 participants using the multivariate classifier (20 with hearing handicap and 90 without).

DT Models

To get a first insight into how the six input features contribute to classification of ears into the two output classes, the rules generated by DT models trained in the 50 iterations were analyzed. As expected, age is a dominant feature within the various splits as it is the feature responsible for the top-level splitting (root node) in all the models. An example of DT model is shown in Figure 2. The figure shows that a cutoff value equal to 52 years can discriminate a subset of 64 out of 138 initial records. Moreover, by further introducing another cutoff value on age at 43 years in the second level of the tree, a sample of 52 no-HL records can be extracted with a Gini index equal to 0 (i.e., pure node). This means that about 70% of the no-HL records in the training set (i.e., 52 out of 74) can be classified by setting a cutoff threshold on age. Another relevant feature for classification is the number of correct responses that is present in the second- and third-level splits, with higher probability of hearing loss associated with a lower number of correct responses (i.e., ≤ 69). A splitting rule involving the percentage of correct responses appears in the fourth-level split, classifying records with a percentage lower than approximately 90% into the HL class. As it can be observed, these cutoffs regarding the most relevant partition rules shown in Figure 2 are similar to those obtained by looking at the discrimination capabilities of the univariate classifiers (see Table 1). A rule associated with SRT appears in the third-level split (i.e., SRT ≤ −16.33 dB SNR); however, this feature seems to be less informative in this model as the rule leads to the identification of only one record in the no-HL class. A splitting rule involving the average reaction time appears only in the fourth-level split, therefore contributing to a limited extent to the classification of ears as it is associated with leaf nodes including a small number of records. Finally, no rules based on total test time are observed. More generally, by looking at the 50 DT models trained on our data set, this feature seems to have

Figure 2. Example of a decision tree from one of the 50 realizations of the training test (Iteration 11). #correct = number of correct responses; %correct = percentage of correct responses; Avg = average; HL = hearing loss.



limited influence on the output class as it tends to appear only in the nodes at the bottom levels.

Post Hoc Explainability Analysis

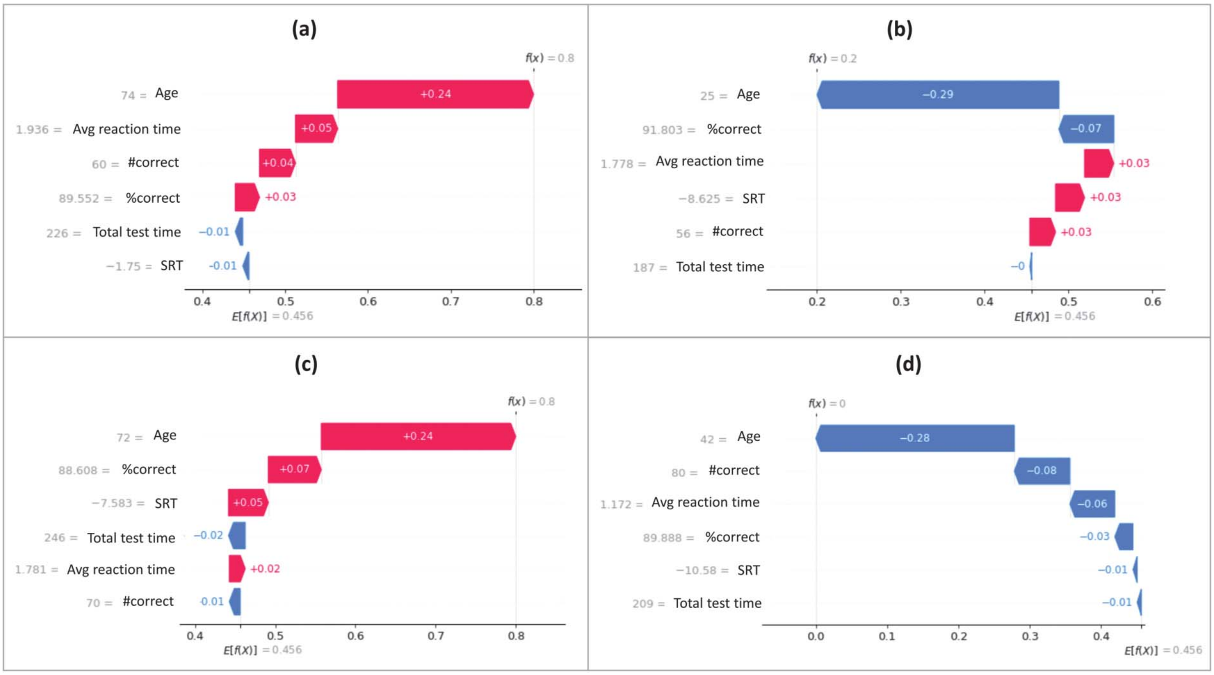
Post hoc explainability techniques have been applied to the models generated using RF, that is, the ML algorithm providing the highest performance measures in identifying HL, as shown in Table 2. Predictions from individual records were evaluated using SHAP values and waterfall visualization, applied to examples of true positive, true negative, false positive, and false negative classifications, to exemplify the role of specific values of features in determining local predictions for each of the two output classes (see the Analysis of Individual Predictions section). Overall predictions were assessed (a) using feature implicit importance and feature permutation importance to assess the role of input features irrespectively of the output class and (b) using PDP visualizations to highlight the role of features and their observed values in determining the probability of classification of records in each of the two output classes (see the Analysis of Overall Predictions section).

Analysis of Individual Predictions

The SHAP technique has been used to investigate the role of specific values of features in determining the output of the model for a certain instance. In fact, for a given feature, different values can contribute to improve or decrease the probability of classification into the HL class. Figure 3 shows the waterfall SHAP diagrams from four different prediction examples (i.e., true positive, true negative, false positive, and false negative) from the test set obtained using one of the 50 RF models generated in this study. For each example in Figure 3, the *y*-axis shows the observed value of each of the input features for the selected record in the data set. For example, Figure 3a shows results from a 74-year-old participant who had an average reaction time of about 1.9 s, a number of correct responses equal to 60, a percentage of correct responses equal to 89.5%, total test time equal to 226 s (i.e., 3 min 46 s), and SRT equal to −1.75 dB SNR. The *x*-axis shows the incremental probability of classification into class HL associated with the observed values of the six input features in the chosen record. The feature values of the true positive example shown in Figure 3a lead to an increase in the probability of classification into the HL class from the initial value of .456 to .8. The initial probability is the prevalence of class HL in the predicted output from the training set in the specific iteration considered in these examples. The color of each bar in the waterfall diagrams indicates the direction of the change in probability of classification into the HL class (positive: red, negative: blue).

From the case represented in Figure 3a, that is, a true positive (i.e., a correctly predicted case of HL), SHAP analysis shows that all features except the total test time and SRT have a positive contribution to the final probability of HL classification, with age (with a value of 74 years) being the most important feature. The average

Figure 3. Analysis of individual predictions using the SHapley Additive exPlanations (SHAP) technique on one of the random forest (RF) models generated in four exemplary records extracted from the data set: (a) true positive, (b) true negative, (c) false positive, and (d) false negative. Avg = average; #correct = number of correct responses; %correct = percentage of correct responses; SRT = speech recognition threshold.



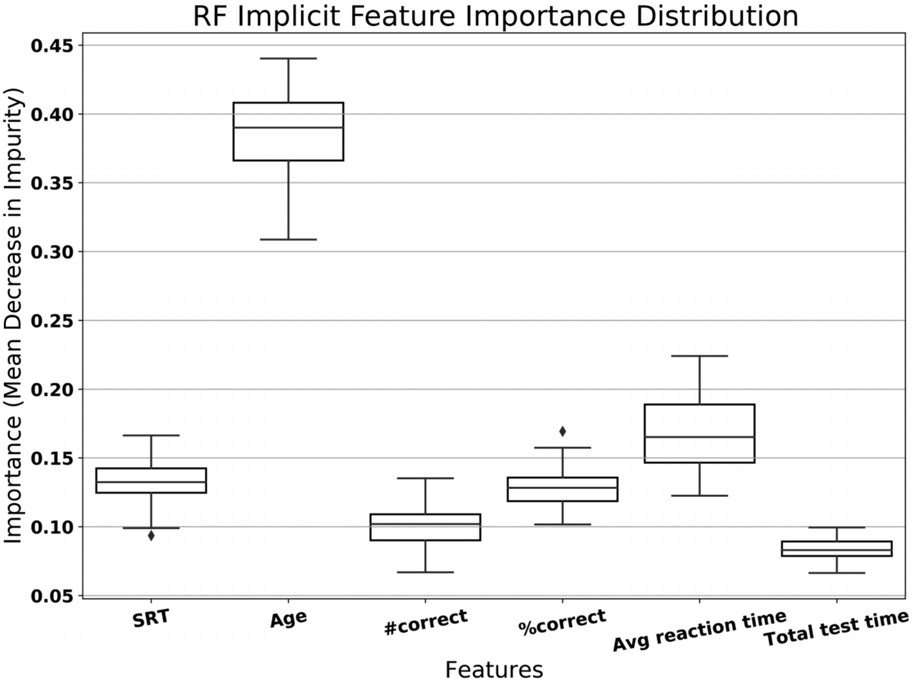
reaction time, the number of correct responses, and the percentage of correct responses show contributions not exceeding 0.05. For the observed values of total test time and SRT, a very small contribution is observed (i.e., a change in probability of about .01). Figure 3b shows a true negative case (i.e., a correctly predicted record from the no-HL class), where the participant's age (i.e., 25 years) is the most relevant feature as it decreases the initial probability of an amount equal to .29, followed by the percentage of correct responses (i.e., a value of about 92% that decreases the probability of classification into the HL class), whereas features such as the average reaction time, SRT, and the number of correct responses had very small contributions (below 0.05), and total test time had zero contribution.

Interestingly, the analysis of SHAP values can also be used to assess the contribution of features when the prediction is not correct, that is, false positives (an incorrectly predicted case of HL such as in Figure 3c) and false negatives (an incorrectly predicted case of no HL such as in Figure 3d), therefore supporting a better understanding of misclassifications. For example, in Figure 3c, the main reason for misclassification of this record into the HL class is the high importance given to the participant's age (i.e., 72 years) when some of the other features have values in line with the average performance of individuals in the no-HL class (e.g., high number and percentage of correct responses, low average reaction time), and SRT is close to the cutoff value identified from univariate analysis (see Table 1). Figure 3d shows an example of a record from the HL class misclassified as no HL where the participant's age is again the main reason for misclassification (i.e., 42 years). The analysis of SHAP values also shows that relatively high number and percentage of correct responses and relatively short average reaction time also contributed, although, to a more limited extent, to misclassification of this record into the no-HL class.

Analysis of Overall Predictions

In addition to the analysis of individual cases shown in the Analysis of Individual Predictions section, general explainability techniques for the analysis of overall classification in the data set might contribute to a better understanding of the overall importance of each feature and the general influence of feature values on the probability of classification in each of the two classes. Figure 4 shows the implicit feature importance distributions, as measured by the mean decrease in impurity, for each of the input features, obtained from RF models trained on 50 realizations of the training set. Age (*Mdn* = 0.386) is notably the most important feature, in line with the results from the DT shown in the DT Models section and in line with the examples of SHAP values shown in the Analysis of Individual Predictions section. The average reaction time (*Mdn* = 0.165), SRT (*Mdn* = 0.131), percentage of correct responses (*Mdn* = 0.129), and number of correct responses

Figure 4. Implicit feature importance of random forest (RF) models obtained from 50 realizations of the training set. SRT = speech recognition threshold; #correct = number of correct responses; %correct = percentage of correct responses; Avg = average.

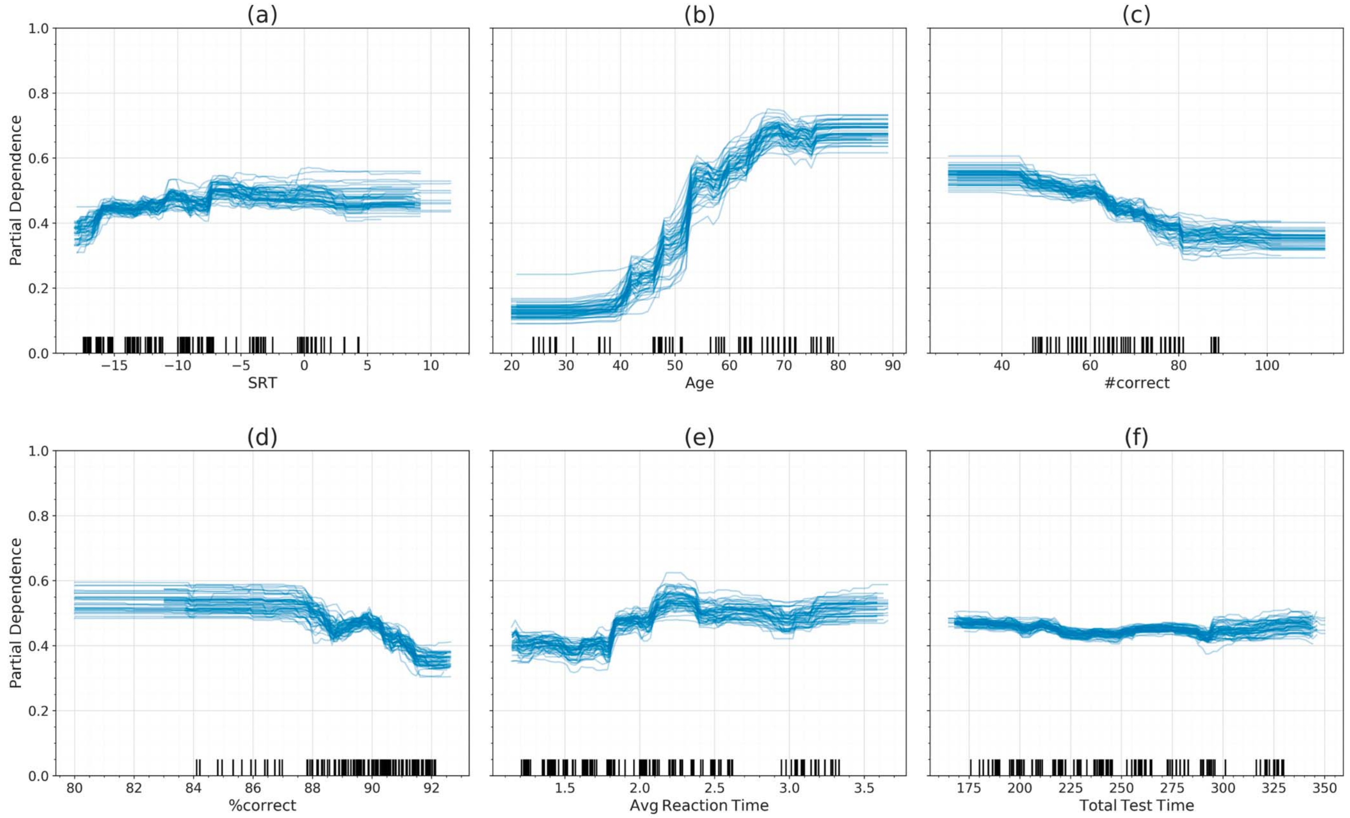


(*Mdn* = 0.104) are relatively close to each other. The less relevant feature is total test time (*Mdn* = 0.082) as the impurity along the RF nodes decreases to an amount below 0.10. Feature permutation technique shows that age (0.218 ± 0.059) is the most important feature since the accuracy of the model decreases and is followed by average reaction time (0.025 ± 0.029). The rest of the features have, on average, very limited importance in comparison to age and average reaction time, namely, total test time (0.007 ± 0.026), number of correct responses (0.005 ± 0.026), percentage of correct responses (0.002 ± 0.029), and SRT (0.000 ± 0.022). These results are remarkably similar to implicit feature importance distributions.

Figure 5 shows the PDP plot of each of the input features computed over 50 realizations of the RF model. In each subpanel, the *y*-axis shows the probability of classification in the HL class and the *x*-axis reflects the distribution of values of the feature across the 50 realizations of the training data set. As in the SHAP examples shown in Figure 3, the PDP values must be interpreted relatively to the proportion of predicted records in the HL class, that is, approximately .5, as averaged across the 50 iterations. Therefore, PDP values above .5 indicate an increasing probability of having hearing loss, whereas PDP values below .5 indicate a decreasing probability of having hearing loss based on the observed data.

Figure 5b shows that age is the feature that exhibits the largest change in probability, from about .1–.2 for individuals younger than 40 years to about .6–.75 for individuals older than about 75 years old. A probability of about .5 is observed at approximately 50–55 years old (i.e., a cutoff value similar to the one shown in Table 1), suggesting that the probability of predicted hearing loss decreases for individuals with age below this age range and that the probability of predicted hearing loss increases for individuals with age above this age range. Other relevant features in terms of probability variation seem to be the number (see Figure 5c) and the percentage of correct responses (see Figure 5d) that vary in a range from about .3–.6, with a probability of about .5 associated with a number of correct responses of approximately 60–65 and a percentage of correct responses from about 88%–90%, and an increasing probability of predicted hearing loss below these cutoff values for both features. Features such as SRT (see Figure 5a) and average reaction time (see Figure 5e) seem to contribute mainly toward the no-HL class as, on average, the observed PDP values tend to be lower than .5 throughout the observed feature ranges. Indeed, the probability of predicted hearing loss decreases below .5 for SRT values approximately lower than −7 dB SNR and for average reaction time values lower than about 2.2 s. However, the range of probability observed for these features is smaller compared to, for example, age, number of correct, and percentage of correct responses. Finally, the total test time shows an almost constant pattern that is associated with the smallest observed change in probability around .5, indicating limited importance

Figure 5. Partial dependence plot (PDP) of the six input features computed from random forest (RF) models obtained from 50 realizations of the training set: (a) speech recognition threshold (SRT), (b) age, (c) number of correct responses (#correct) responses, (d) percentage of correct responses (%correct), (e) average (Avg) reaction time, and (f) total test time.



of this feature in the model's output, in line with the results shown in Figure 4.

Analysis of Predictions in Younger Versus Older Participants

To better emphasize the role of features derived from the speech-in-noise test in predicting the hearing loss class, the training set was partitioned into two subsets using the age cutoff value identified with univariate analysis (i.e., 53 years), and DTs were trained separately on each of these subpopulations (see Figure 6). Figure 6a shows the splitting rules of an exemplary DT generated using all features, except age, on subjects aged 53 years or younger. The first-, second-, and fourth-level nodes involve the percentage and the number of correct responses; however, only a small number of records is discriminated by these rules (e.g., one record is classified as HL if percentage of correct responses is > 92.82, and another record is classified as HL if number of correct responses is ≤ 45). Vice versa, the rule at the third-level split has higher classification ability as an SRT cutoff around −10 dB SNR can correctly identify 44 records as no HL, with no false positives.

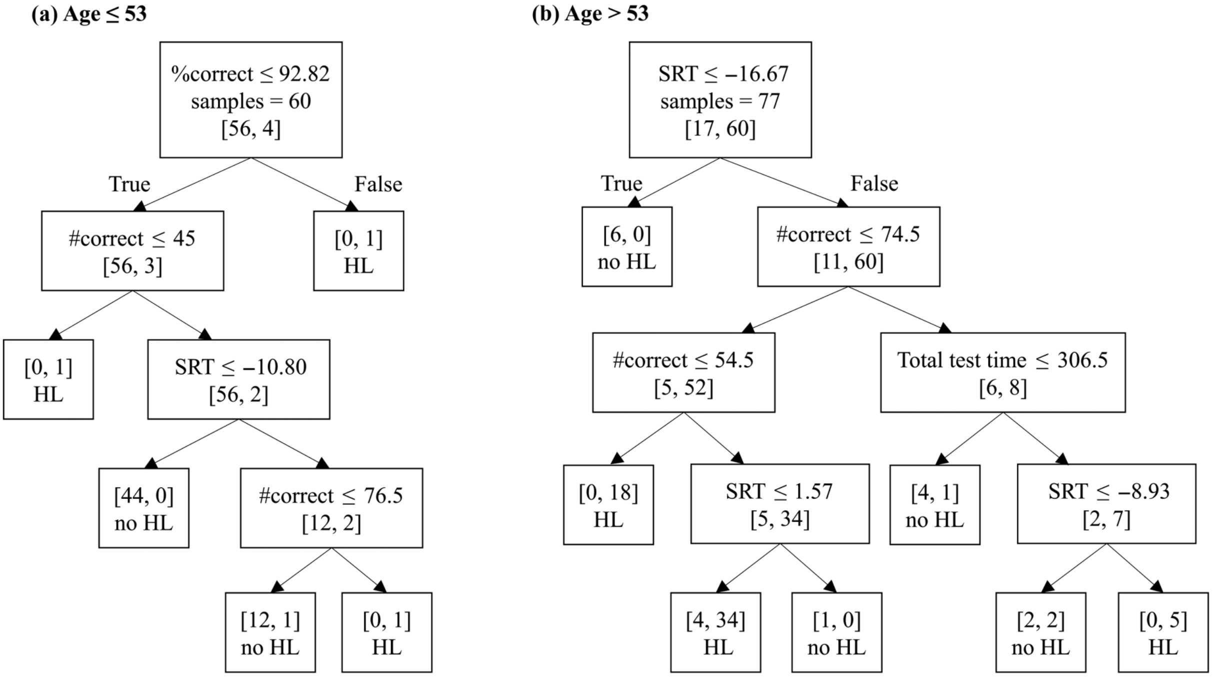
Figure 6b shows the splitting rules of an example of DT trained on the subset of records extracted from subjects older than 53 years. The root node allows to correctly discriminate records in the no-HL class based on very low values of SRT, that is, below −16.67 dB SNR. Moreover, records with SRT above this cutoff but a small number of correct responses (i.e., 74 or lower) are likely to be classified in the HL class. Indeed, 52 out of 60 subjects with hearing loss and older than 53 years of age achieved less than 54 correct responses. In case of a number of correct responses higher than 74 and a total test time higher than 5 min, the SRT is again the discriminating feature, as subjects with SRTs greater than −9 dB SNR are associated with hearing loss.

Discussion

Univariate Classification Performance

Univariate classification performance was used as a benchmark to compare the individual discrimination capability of each of the six input features (see Table 1). The

Figure 6. Examples of decision trees trained on two subpopulations, defined by the age cutoff shown in Table 1: (a) age ≤ 53 years and (b) age > 53 years. %correct = percentage of correct responses; #correct = number of correct responses; HL = hearing loss; SRT = speech recognition threshold.



analysis of univariate classification performance suggests that most of the features considered in this study are potentially relevant for identifying hearing loss. Indeed, five out of the six features considered exhibit an AUC > .7, and four out of six exhibit accuracy > .7. Age shows the highest association with hearing loss, and a cutoff value of about 53 years between the HL and no-HL classes is identified from the univariate classification, in line with evidence from the literature. Increasing hearing thresholds are part of the gradual decline of hearing capabilities that often becomes a concern starting around 60–65 years and gets progressively worse (Purnami et al., 2020). Recently, using the higher WHO cutoff for defining hearing loss (i.e., PTA > 25 dB HL), we found that age was a strong predictor of hearing loss (Polo, Zanet, Paglialonga, & Barbieri, 2021; Zanet et al., 2021). According to this first univariate analysis, age alone would seem to be a valid predictor of mild hearing loss, in line with the known evidence that age is the strongest predictor of hearing loss among adults aged 20–69 years, with the greatest amount of hearing loss in the 60–69 years age group (Hoffman et al., 2017). However, the specificity of age alone is below 70%. Thus, considering exclusively age as the discriminating variable can lead to a high rate of normal-hearing ears misclassified as having hearing loss. As shown in Table 1, the use of a pool of variables can improve the discrimination of the two classes of subjects, reaching an accuracy that is just below 90%. Specifically, features such as SRT and the number of correct responses seem to increase the capability to identify the normal-hearing class (specificity = 81%).

Some of the features extracted from the speech-in-noise test are inherently related to each other as they reflect the individual performance on a given speech-in-noise task. For example, the Spearman's correlation between the number and the percentage of correct responses is .53, and that between the average reaction time and SRT is equal to .49, in line with neurophysiological evidence that the same underlying neural mechanism determine perceptual decisions, confidence, and reaction time (Fetsch et al., 2014). Moreover, age is correlated with speech recognition performance (e.g., the correlation between age and percentage of correct responses is equal to −.57) as well as with the average reaction time (correlation = .66), in line with the study by Ratcliff et al. (2001). However, all the observed correlations were in the low-to-moderate range (.30–.70), and none was in the high or very high range (i.e., above .70; Mukaka, 2012).

Although the SRT obtained from speech-in-noise tests and hearing thresholds measured using pure-tone audiometry represent two inherently different aspects of hearing ability, a moderate correlation has been reported between these two measures (Bosman, & Smoorenburg, 1995; Leensen et al., 2011; Smoorenburg, 1992). For example, Leensen et al. (2011) found a correlation of .66, .69, and .72 between pure-tone hearing thresholds at 0.5, 1, 2, and 4 kHz and SRTs extracted from three different speech-in-noise tests, the Dutch version of the digit triplet test, Earcheck, and Occupational Earcheck, respectively, in 98 subjects, half of whom had different degrees of noise-induced hearing loss. Decreased consonant recognition is one of the first signs of age-related hearing loss (Killion & Niquette, 2000), and lower speech recognition abilities with age have been widely demonstrated in the literature (e.g., Heidari et al., 2018). The correlation between SRT and pure-tone hearing thresholds, as derived from this study, is equal to .63. However, other features, in addition to SRT, should be considered as predictors of hearing loss in order to obtain more accurate detections. The analysis of univariate classification performance also showed that the total test time alone seems not to be a significant predictor of the output class because of a kind of compensation effect, whereas the average reaction time appears to be more meaningful. Specifically, individuals with no hearing loss tend to require a higher number of trials before finishing the test as they can reach very low SNRs in the adaptive procedure, but they tend to be quicker in recognizing each single stimulus (i.e., lower average reaction time). On the other hand, individuals with hearing loss tend to require a lower number of trials as they tend to reach the stopping criterion earlier due to a higher number of errors in the adaptive procedure, but, on average, they tend to have a higher average reaction time. As a result, the total test time remains substantially similar in individuals with and without hearing loss, as confirmed by the minor role of total test time in predicting hearing loss.

In general, the cutoff values obtained on the individual features are very similar to those previously obtained on a smaller sample of 156 records using the higher cutoff in the WHO definition of hearing loss (PTA > 25 dB HL; Zanet et al., 2021). Only the age cutoff decreased from 64 to 53 years, presumably because of the lowering of the cutoff PTA value; therefore, a higher proportion of ears with hearing loss is observed in younger subjects.

Multivariate Classification Performance

Figure 2 shows that some of the cutoffs in the most significant DT rules (i.e., age = 52 years, number of correct responses = 69.50, percentage of correct responses = 90.37%) are very close to the cutoff values shown in Table 1 (age = 53 years, number of correct responses = 63, percentage of correct responses = 90.28%). The cutoff values obtained by considering single predictors are therefore maintained when generalizing to a multivariate DT approach.

When considering different ML algorithms, similar average performance is observed, as determined by running 50 iterations on different realizations of the training and test data sets (see Table 2). Although being transparent and therefore able to provide all the information about the mechanisms that lead to a certain prediction through a set of explicit rules, DTs have the lowest performance among the seven algorithms tested, whereas RF, SVM, LR, and GB had significantly higher performance, with RF having the best trade-off between sensitivity (.86) and specificity (.85). In addition to the observed performance in terms of classification accuracy, sensitivity, and specificity, the proposed system also provides consistent identification of individuals in relation to their self-perceived hearing handicap, as measured by the HHIE-S questionnaire score (see Figure 1). Specifically, the predictions of HL class generated using both the measured SRT and the multivariate classification approach are related to a higher level of self-perceived hearing handicap compared to the no-HL class, and the multivariate classifier can correctly classify a higher number of participants into the correct self-reported hearing handicap class compared to the univariate one. Moreover, significantly higher HHIE-S scores are associated with higher SRT, indicating a decreased ability to recognize speech in a background noise that may be associated with higher self-perceived hearing handicap.

In general, the performance obtained using multivariate ML models shows similar or even better results with respect to the classification performance of other speech-in-noise tests based on SRT only. For example, the U.S. version of the digits-in-noise test using the same criterion for the identification of hearing loss (i.e., 20 dB HL) and a cutoff for SRT equal to −5.7 dB SNR gave a sensitivity of .80 and a specificity of .83 (Watson et al., 2012). Moreover, the classification performance obtained in this study is slightly better with respect to our previous study where sensitivity and specificity were around .79 on a sample of 156 ears (Zanet et al., 2021). These differences may be due to a combination of factors such as the use of a different criterion for defining hearing loss in this study compared to our earlier investigations (i.e., 20 dB HL vs. 25 dB HL, respectively). Also, the presence of a more balanced data set in this study compared to our earlier study (i.e., 46% vs. 34% of records in the HL class) may have influenced our results, as in this study, models have been more equally trained on the two classes. The performance of the proposed models in identifying hearing loss can be considered satisfactory, and the use of post hoc explainability techniques provides additional insights into the models' prediction mechanisms. However, in real-world settings, predictive models are not applied on class-balanced data, and therefore, the expected number of false positives and false negatives will be different. For example, we can consider the average confusion matrix on the test set obtained from the best performing model (RF) across the 50 iterations (i.e., false negative rate = 15%, false positive rate = 17.4%, accuracy = 83.7%, and negative predictive value = 86% on a data set including 46% of HL records) and extrapolate the performance based on the assumption that sensitivity and specificity are kept constant. Assuming a prevalence of hearing loss of about 20% in adults (e.g., Stevens et al., 2013) and given a hypothetical sample of 100 cases, with similar characteristics to those of the general adult population (i.e., 20 in the HL class and 80 in the no-HL class), the system would predict about 69 cases as no HL (i.e., 0.15 × 20 + (1 − 0.174) × 80), of which only three would be false negatives and 66 (i.e., 96%) would be actual normal-hearing cases (true negatives). Therefore, the negative predictive value of the system would increase from about 86% to about 96%, and the accuracy would slightly decrease from about 83.7% to about 83% (i.e., 66 true negatives + 17 true positives). Specifically, the ability of the system to provide a correct no-HL decision to cases with no hearing loss would improve when the results are extrapolated to the real-world prevalence of hearing loss as the percentage of cases with hearing loss classified as no HL by the system would decrease to only 4%, therefore potentially reducing the number of individuals with hearing loss who should be recommended follow-up examinations but are not identified during the screening.

Post Hoc Explainability of Input Features

Post hoc explainability techniques have been applied to RF, that is, the algorithm that reached the highest classification performance (see Table 2), to analyze the importance of each of the input features in determining individual and overall predictions. The analysis of SHAP values in individual cases (see Figure 3) showed that, in general, the individual's age was the most important feature in terms of probability of classification into a given output class as it led to an increase/decrease of the probability of classification into the HL class of an amount between approximately 0.2 and 0.3. The importance of age in the trained models is related to the statistical distribution of data in our data set as individuals with hearing loss were, in general, older than those with normal hearing, in line with the typical distribution of age-related hearing loss in the general population (U.S. Preventive Services Task Force et al., 2021). The results obtained using other post hoc techniques such as implicit feature importance (see Figure 4), feature permutation importance, and PDP (see Figure 5) on overall predictions confirm that age is the most relevant feature, in line with results obtained from univariate analysis and from DT rules. The results obtained from feature implicit importance and feature permutation importance indicate average reaction time as the second most relevant feature and very low importance of the rest of the features. However, these techniques give only absolute values of the importance of a given feature irrespectively of the output class as they do not provide specific information about the direction of the contribution of the feature toward the HL or the no-HL class in the model prediction. Therefore, if a given feature is important for predicting only one of the two classes, it may show limited overall importance when implicit feature importance or feature permutation importance techniques are used. On the other hand, information about the direction of a feature's contribution is provided by PDP analysis, which also shows how the probability of a given output class in the model prediction is related to the distribution of each feature's values (see Figure 5). Specifically, PDP analysis indicates a cutoff of about 50–55 years for age, whereby higher values contribute toward the HL class and lower values toward the no-HL class. Moreover, since the PDP plot of age shows the biggest variability in probability, it further confirms that age is the most important feature to predict the output class in the data set used here. Similarly, the PDP indicated that both the number and the percentage of correct responses contributed to each of the two output classes, although the observed changes in probabilities are smaller. Some features contributed mainly to classification into the no-HL class, specifically average reaction time and SRT, but their contribution was smaller compared to the other features. It is worth noting that, since the results of 50 iterations are pooled in the PDP in Figure 5, the observed cutoff values are approximated. Nevertheless, the cutoff values shown in the PDP are similar to the ones provided by the DT rules (see Figure 1), hence supporting the use of RF in combination with post hoc explainability techniques as a means to achieve, concurrently, accuracy of classification and explainability of predictions.

Age Contribution

As previously discussed, our multivariate analysis shows that age has a dominant contribution in discriminating normal-hearing ears from ears with hearing loss. However, other variables in addition to age were found to be important for this purpose. Specifically, Table 1 shows that features such as average reaction time, SRT, percentage of correct responses, and number of correct responses have accuracy around .7 or higher and that the accuracy obtained using all the features is higher than that obtained using age alone. These findings are further confirmed by the analysis of feature and value of importance shown in Figures 5 and 6. The bias introduced by age in the classification models developed here is also demonstrated by the analysis of exemplary misclassifications shown in Figure 3, whereby the false positives and false negatives cases are mainly determined by the values of age.

To analyze in more detail the role of other features derived from the speech-in-noise test, we analyzed DT models trained on two sub–data sets, specifically from individuals younger than or older than 53 years of age (see Figure 6). The rationale was to identify two subpopulations that show a different risk of presenting hearing loss due to their age as older individuals are more likely to have hearing loss (Haile et al., 2021). As shown in Figure 6a, the SRT would seem to discriminate between the two classes in younger individuals, as ears with an SRT lower than a cutoff of about −10 dB SNR are more likely to be normal hearing. Similarly, SRT is also determinant in classifying subjects who seem to be at higher risk of suffering from hearing loss because of their age (see Figure 6b) but who actually show very good speech recognition in noise (e.g., highly negative SRT, below −16.67 dB SNR, or high number of correct responses, above 74). Hence, the analysis of features extracted from the speech-in-noise test, particularly the SRT and the number of correct responses, might be helpful to identify hearing loss in individuals at higher or lower risk of developing hearing loss.

Limitations and Future Developments

This study demonstrated that multivariate ML models based on features extracted from a speech-in-noise test may be a promising tool to accurately identify hearing loss. However, it is acknowledged that this study has some limitations. The use of speech stimuli from an English speaker and the related transcription in Roman alphabet is supported by our earlier estimates of VCV recognition performance across five languages from computational simulations and listening tests (Rocco, 2018) and by the fact that English is the top language by total number of speakers worldwide and the most widely used language in the web (Eberhard et al., 2019; Internet World Stats, 2020). As such, participants are likely to have had previous experience with spoken and written English. Several studies have investigated the effect of native language on speech-in-noise recognition performance (e.g., Lecumberri et al., 2010). In the area of speech-in-noise screening tests, some studies showed that nonnative subjects might present poorer recognition performance with respect to native ones with the digits-in-noise hearing tests, that is, a test using language-specific speech stimuli (e.g., Potgieter et al., 2018; Taylor et al., 2020). However, in our sample, we tested using meaningless intervocalic consonants, and we did not observe differences in the distributions of SRT and age as a function of the output class (HL vs. no HL) between native and nonnative English listeners, in line with preliminary results from our earlier study (Paglialonga et al., 2020). Moreover, the distributions of SRT in different nonnative English listeners (i.e., Italian vs. other languages, 19 age-matched subjects in each subgroup) were similar. However, these findings are preliminary, and further validation of the robustness of the test results across different languages is needed. It will be important in future studies to assess the effects of native language and address test performance in listeners of characters-based and non-Roman alphabet languages.

Trusting an ML decision depends not only on the interpretability of the model but also on the quality of the data used for training (i.e., its completeness and its ability to depict a wide range of possible situations; Rudin, 2019). As it is well known, ML models are data-driven, and in order to limit bias, it is necessary that the sample used for training the ML models has an adequate size and that it reflects as much as possible the reality of the phenomenon to be modeled. Moreover, applying standardization techniques on a small data set may introduce some sort of bias, as the training set may not perfectly represent the original phenomenon. This issue could be partially addressed by normalizing the data using normative values. For the sake of expanding the size of the data set, techniques such as data augmentation can be helpful as they can generate large high-quality, balanced data sets from relatively small ones using, for example, generative adversarial networks (Tran et al., 2021; Vaccari et al., 2021). On the other hand, to overcome the issues related to bias in data collection, computational techniques are of limited help, and further experimental data are needed as the specific experiment settings that led to the development of our data set might have influenced the results in terms of feature importance. In our data set, most of the records in the HL class were from older subjects, and most of the records in the no-HL class were from younger subjects, as usually observed in the general population. As a result, the ML models trained on this data set use age as the most important predictor of hearing loss, and therefore, older individuals are more likely to be classified in the HL class and younger individuals are more likely to be classified in the no-HL class (e.g., misclassification examples in Figures 3c and 3d). The analysis of DTs in younger versus older individuals (see Figure 6) suggests that features extracted from the speech-in-noise test, particularly the SRT and the number of correct responses, may help in identifying hearing loss in individuals at higher or lower risk of hearing loss (i.e., older vs. younger individuals, respectively). These models are trained on small data sets, and it will be important to analyze the full range of rules and cutoff values determined on a larger, more age-balanced data set. Specifically, the data used here were mainly collected during hearing loss prevention initiatives that were primarily addressed to older adults. As such, recruitment largely involved participants who had already elected to attend a hearing screening and who were likely to have a hearing problem. In fact, the distribution of age and hearing loss in the study sample are skewed toward the older age groups, and the observed prevalence of hearing loss is higher than that reported in the general population. For example, in our sample, 31% of participants have an age of ≥ 65 years old, compared to an average of 20.8% in the European Union and 23.3% in Italy (World Bank Open Data, n.d.), where most of our participants come from. As such, the distribution of age in the study sample may have led to classification bias. Although the results of this study are in line with knowledge from the literature (i.e., higher prevalence of hearing loss in older adults), the classification models trained here may not fully reflect the role of each of the input features in determining the likelihood of classification in the HL class. Future research on a representative data set, with a higher proportion of young and middle-age adults, will be useful to optimize the proposed ML approach for application to the general population.

The data set used here presents a certain degree of multicollinearity among predictors, with the highest correlation between average reaction time and age (i.e., .66). Multicollinearity does not, in general, affect predictions of multivariate models (Kutner et al., 2005), but it might affect model interpretation and the explainability analysis. XAI techniques tell us how much a feature is important in the model logic, but this is not necessarily coincident with how much the feature is important in the real world as the importance of certain features might become less evident due to multicollinearity and redundancy. The strength of this analysis may be improved by considering a broader set of meaningful and independent features. Additional features might be extracted from the speech-in-noise test, for example, features that measure recognition performance separately for low- and high-frequency consonants, measures of reaction time as a function of the answer (correct/wrong), type of stimulus (low/high frequency), or SNR. Ongoing research focuses on the development and evaluation of a web-based platform for the collection of a large set of features, including risk factors (Paglialonga et al., 2022).

This study investigates ML approaches to identify hearing loss based on features extracted from a specific speech-in-noise test. In future studies, it will also be interesting to evaluate features extracted from other types of validated speech-in-noise tests that are widely available in multiple languages (e.g., the digit triplet test) in order to investigate whether the ability of ML algorithms to predict hearing loss is affected by the type of underlying speech-in-noise test. Moreover, in this study, only a binary variable derived from pure-tone testing has been used to determine the output class. However, pure-tone thresholds give only a partial picture of the real-life hearing and communication abilities, and the screening outcome addressed in this study (HL vs. no HL) does not consider important aspects related to perceived handicap or treatment benefit. Future studies will be needed to evaluate the classification performance of the proposed predictive approach to assess different output variables such as the degree of hearing loss as assessed in follow-up visits, the self-perceived hearing handicap or hearing disability, or the measured hearing amplification benefit.

More generally, a multivariate, explainable approach such as the one developed here could be generalized and applied in different audiological applications to analyze the role of multiple variables and factors in determining a certain condition. For example, multivariate approaches could help address the factors leading to cognitive decline in older adults. Indeed, multiple measures from a battery of tests (e.g., pure-tone thresholds, cognitive measures from questionnaires of functional tests, self-assessment questionnaires of hearing handicap, electrophysiological measures, medical history, or even demographic and socioeconomic indicators) could be combined in future studies to gain a deeper knowledge about the factors that might lead to cognitive decline in individuals with hearing loss.

Conclusions

This study introduced a multivariate framework based on ML algorithms and explainability techniques to investigate the predictive capabilities of features extracted from a speech-in-noise test for the sake of identifying hearing loss in adults. The results of this study indicated, for the first time, that a multivariate approach using features such as the subject's age, the number and percentage of correct responses, and the average reaction time, in addition to SRT, may play a role in identifying hearing loss in adults. The observed classification performance is high (sensitivity = .86, specificity = .85) in the balanced data set used here. When the observed performance is used to estimate the expected performance of the system in a hypothetical population with hearing loss prevalence similar to that commonly observed in real-world settings, very high negative predictive value (i.e., 96%) and high accuracy (i.e., 83%) are estimated. However, it will be essential to further expand the validation of the proposed multivariate algorithms for hearing loss identification on a larger representative population to fully demonstrate their performance in real-world settings. For example, it will be important to involve more participants with varying degrees of hearing loss across the entire age range and across a broader spectrum of native languages, together with the investigation of other validated speech-in-noise tests. Lastly, it will be interesting to assess whether expanded sets of input features are more accurate in predicting hearing loss than smaller sets such as the one investigated here by using additional ML algorithms and explainability techniques to better highlight the role of features in determining prediction and ultimately demonstrate the viability of speech-in-noise tests for identifying hearing loss for the sake of adult hearing screening.

Data Availability Statement

Codes and sample data are available at https://github.com/lenattimarta/whisper_posthocXAI

Author Contributions

Marta Lenatti: Methodology (Lead), Software (Lead), Formal analysis (Lead), Validation, Writing – original draft (Lead), Writing – review & editing (Lead). **Pedro A. Moreno-Sánchez:** Methodology (Lead), Software (Lead), Formal analysis (Lead), Writing – original draft (Supporting), Writing – review & editing (Supporting). **Edoardo M. Polo:** Investigation (Lead), Software (Lead), Formal analysis (Supporting), Writing – review & editing (Supporting). **Maximiliano Mollura:** Software (Supporting), Writing – review & editing (Supporting). **Riccardo Barbieri:** Conceptualization (Supporting), Funding acquisition (Supporting), Supervision (Supporting), Writing – review & editing (Supporting). **Alessia Paglialonga:** Methodology (Supporting), Investigation (Supporting), Conceptualization (Lead), Funding acquisition (Lead), Supervision (Lead), Project administration, Writing – original draft (Lead), Writing – review & editing (Lead).

Acknowledgments

This research work was partially supported by Capita Foundation (Project WHISPER, Widespread Hearing Impairment Screening and PrEvention of Risk, 2020 Auditory Research Grant). The authors would like to thank the Lions Clubs International and Associazione La Rotonda, Baranzate (Milan), for their contribution in the organization and management of experiments in unscreened adults. The authors are also grateful to Anna Bersani, Carola Butera, Antonio Carrella, and Marco Zanet, who contributed to data collection on unscreened adults.

References

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data, 4,* 688969. https://doi.org/10.3389/fdata.2021.688969

Blamey, P. J., Blamey, J. K., & Saunders, E. (2015). Effectiveness of a teleaudiology approach to hearing aid fitting. *Journal of Telemedicine and Telecare, 21*(8), 474–478. https://doi.org/10.1177/1357633x15611568

Bosman, A. J., & Smoorenburg, G. F. (1995). Intelligibility of Dutch CVC syllables and sentences for listeners with normal-hearing and with three types of hearing impairment. *Audiology, 34*(5), 260–284. https://doi.org/10.3109/00206099509071918

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Cooke, M., Lecumberri, M. L. G., Scharenborg, O., & van Dommelen, W. A. (2010). Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages. *Speech Communication, 52*(11–12), 954–967. https://doi.org/10.1016/j.specom.2010.04.004

Dalton, D. S., Cruickshanks, K. J., Klein, B. E., Klein, R., Wiley, T. L., & Nondahl, D. M. (2003). The impact of hearing loss on quality of life in older adults. *The Gerontologist, 43*(5), 661–668. https://doi.org/10.1093/geront/43.5.661

Davies, H. R., Cadar, D., Herbert, A., Orrell, M., & Steptoe, A. (2017). Hearing impairment and incident dementia: Findings from the English longitudinal study of ageing. *Journal of the American Geriatrics Society, 65*(9), 2074–2081. https://doi.org/10.1111/jgs.14986

Davis, A., Smith, P., Ferguson, M., Stephens, D., & Gianopoulos, I. (2007). Acceptability, benefit and costs of early screening for hearing disability: A study of potential screening tests and models. *Health Technology Assessment, 11*(42), 1–294. https://doi.org/10.3310/hta11420

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2019). *Ethnologue: Languages of the world* (22nd ed.). http://www.ethnologue.com

Feltner, C., Wallace, I. F., Kistler, C. E., Coker-Schwimmer, M., & Jonas, D. E. (2021). Screening for hearing loss in older adults. *JAMA, 325*(12), 1202. https://doi.org/10.1001/jama.2020.24855

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron, 83*(4), 797–804. https://doi.org/10.1016/j.neuron.2014.07.011

Fisher, A., Rudin, C., & Dominici, F. (2019). *All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously*. ArXiv. http://arxiv.org/abs/1801.01489

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Haile, L. M., Kamenov, K., Briant, P. S., Orji, A. U., Steinmetz, J. D., Abdoli, A., Abdollahi, M., Abu-Gharbieh, E., Afshin, A., Ahmed, H., Ahmed Rashid, T., Akalu, Y., Alahdab, F., Alanezi, F. M., Alanzi, T. M., Al Hamad, H., Ali, L., Alipour, V., Al-Raddadi, R. M., . . . Chadha, S. (2021). Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the Global Burden of Disease Study 2019. *The Lancet, 397*(10278), 996–1009. https://doi.org/10.1016/s0140-6736(21)00516-x

Heidari, A., Moossavi, A., Yadegari, F., Bakhshi, E., & Ahadi, M. (2018). Effects of age on speech-in-noise identification: Subjective ratings of hearing difficulties and encoding of fundamental frequency in older adults. *Journal of Audiology and Otology, 22*(3), 134–139. https://doi.org/10.7874/jao.2017.00304

Hoffman, H. J., Dobie, R. A., Losonczy, K. G., Themann, C. L., & Flamme, G. A. (2017). Declining prevalence of hearing loss in U.S. adults aged 20 to 69 years. *JAMA Otolaryngology—Head & Neck Surgery, 143*(3), 274–285. https://doi.org/10.1001/jamaoto.2016.3527

Humes, L. E. (2013). Understanding the speech-understanding problems of older adults. *American Journal of Audiology, 22*(2), 303–305. https://doi.org/10.1044/1059-0889(2013/12-0066)

Internet World Stats. (2020). *Internet world users by language: Top 10 languages*. https://www.internetworldstats.com/stats7.htm

Killion, M. C., & Niquette, P. A. (2000). What can the pure-tone audiogram tell us about a patient's SNR loss? *The Hearing Journal, 53*(3), 46–48. https://doi.org/10.1097/00025572-200003000-00006

Kumar, U. (2019). Applications of machine learning in disease pre-screening. In T. Edoh, P. Pawar, & S. Mohammad (Eds.), *Pre-screening systems for early disease prediction, detection, and prevention* (pp. 278–320). IGI Global. https://doi.org/10.4018/978-1-5225-7131-5.ch010

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Lecumberri, M. R., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Leensen, M. C., de Laat, J. A., & Dreschler, W. A. (2011). Speech-in-noise screening tests by Internet, Part 1: Test evaluation for noise-induced hearing loss identification. *International Journal of Audiology, 50*(11), 823–834. https://doi.org/10.3109/14992027.2011.595016

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Luor, D. C. (2015). A comparative assessment of data standardization on support vector machine for classification problems. *Intelligent Data Analysis, 19*(3), 529–546. https://doi.org/10.3233/IDA-150730

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59*(3), 203–243. https://doi.org/10.1016/j.cogpsych.2009.04.001

Moreno-Sanchez, P. A. (2020). Development of an explainable prediction model of heart failure survival by using ensemble trees. In X. Wu, C. Jermaine, L. Xiong, X. Hu, O. Kotevska, S. Lu, W. Xu, S. Aluru, C. Zhai, E. Al-Masri, Z. Chen, & J. Saltz (Eds.), *2020 IEEE International Conference on Big Data* (pp. 4902–4910). https://doi.org/10.1109/BigData50022.2020.9378460

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71.

Nuesse, T., Steenken, R., Neher, T., & Holube, I. (2018). Exploring the link between cognitive abilities and speech recognition in the elderly under different listening conditions. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00678

Paglialonga, A., Grandori, F., & Tognola, G. (2013). Using the Speech Understanding in Noise (SUN) test for adult hearing screening. *American Journal of Audiology, 22*(1), 171–174. https://doi.org/10.1044/1059-0889(2012/12-0055)

Paglialonga, A., Lenatti, M., Polo, E. M., Paolini, M., Petrella, L., Mollura, M., & Barbieri, R. (2022). *WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk): A new platform for early identification of hearing impairment and cognitive decline*. Hearing Across the Lifespan Conference (HEAL), June 16–18, 2022, Cernobbio, Italy.

Paglialonga, A., Polo, E. M., Zanet, M., Rocco, G., van Waterschoot, T., & Barbieri, R. (2020). An automated speech-in-noise test for remote testing: Development and preliminary evaluation. *American Journal of Audiology, 29*(3S), 564–576. https://doi.org/10.1044/2020_aja-19-00071

Paglialonga, A., Tognola, G., & Grandori, F. (2014). A user-operated test of suprathreshold acuity in noise for adult hearing screening: The SUN (Speech Understanding in Noise) test. *Computers in Biology and Medicine, 52*, 66–72. https://doi.org/10.1016/j.compbiomed.2014.06.012

Polo, E. M., Zanet, M., Lenatti, M., van Waterschoot, T., Barbieri, R., & Paglialonga, A. (2021). Development and evaluation of a novel method for adult hearing screening: Towards a dedicated smartphone app. In R. Goleva, N. R. da Cruz Garcia, & I. M. Pires (Eds.), *IoT Technologies for HealthCare: 7th EAI International Conference, HealthyIoT 2020, Viana do Castelo, Portugal, December 3, 2020, Proceedings* (pp. 3–19). Springer. https://doi.org/10.1007/978-3-030-69963-5_1

Polo, E. M., Zanet, M., Paglialonga, A., & Barbieri, R. (2021). Preliminary evaluation of a novel language independent speech-in-noise test for adult hearing screening. In T. Jarm, A. Cvetkoska, S. Mahnič-Kalamiza, & D. Miklavcic (Eds.), *8th European Medical and Biological Engineering Conference* (pp. 976–983). Springer. https://doi.org/10.1007/978-3-030-64610-3_109

Potgieter, J. M., Swanepoel, W., Myburgh, H. C., & Smits, C. (2018). The South African English smartphone digits-in-noise hearing test: Effect of age, hearing loss, and speaking competence. *Ear and Hearing, 39*(4), 656–663. https://doi.org/10.1097/AUD.0000000000000522

Purnami, N., Mulyaningsih, E. F., Ahadiah, T. H., Utomo, B., & Smith, A. (2020). Score of Hearing Handicap Inventory for the Elderly (HHIE) compared to Whisper Test on presbycusis. *Indian Journal of Otolaryngology and Head & Neck Surgery*. https://doi.org/10.1007/s12070-020-01997-5

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16*(2), 323–341. https://doi.org/10.1037/0882-7974.16.2.323

Reed, N. S., Altan, A., Deal, J. A., Yeh, C., Kravetz, A. D., Wallhagen, M., & Lin, F. R. (2019). Trends in health care costs and utilization associated with untreated hearing loss over 10 years. *JAMA Otolaryngology—Head & Neck Surgery, 145*(1), 27–34. https://doi.org/10.1001/jamaoto.2018.2875

Rocco, G. (2018). *Design, implementation, and pilot testing of a language-independent speech intelligibility test* [Master's thesis, ING–Scuola di Ingegneria Industriale e dell'Informazione/School of Industrial and Information Engineering]. Politecnico di Milano. https://www.politesi.polimi.it/handle/10589/140319

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery, 8*(4), e1249. https://doi.org/10.1002/widm.1249

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology, 43*(1), 15–28. https://doi.org/10.1080/14992020400050004

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America, 91*(1), 421–437. https://doi.org/10.1121/1.402729

Stevens, G., Flaxman, S., Brunskill, E., Mascarenhas, M., Mathers, C. D., Finucane, M., & Global Burden of Disease Hearing Loss Expert Group. (2013). Global and regional hearing impairment prevalence: An analysis of 42 studies in 29 countries. *European Journal of Public Health, 23*(1), 146–152. https://doi.org/10.1093/eurpub/ckr176

Taylor, H., Shryane, N., Kapadia, D., Dawes, P., & Norman, P. (2020). Understanding ethnic inequalities in hearing health in the U.K.: A cross-sectional study of the link between language proficiency and performance on the Digit Triplet Test. *BMJ Open, 10*(12), e042571. https://doi.org/10.1136/bmjopen-2020-042571

Tran, N.-T., Tran, V.-H., Nguyen, N.-B., Nguyen, T.-K., & Cheung, N.-M. (2021). On data augmentation for GAN training. *IEEE Transactions on Image Processing, 30,* 1882–1897. https://doi.org/10.1109/tip.2021.3049346

U.S. Preventive Services Task Force., Krist, A. H., Davidson, K. W., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., Donahue, K. E., Doubeni, C. A., Epling, J. W., Jr., Kubik, M., Li, L., Ogedegbe, G., Pbert, L., Silverstein, M., Stevermer, J., Tseng, C. W., & Wong, J. B. (2021). Screening for hearing loss in older adults: U.S. Preventive Services Task Force recommendation statement. *JAMA, 325*(12), 1196–1201. https://doi.org/10.1001/jama.2021.2566

Vaccari, I., Orani, V., Paglialonga, A., Cambiaso, E., & Mongelli, M. (2021). A Generative Adversarial Network (GAN). *Technique for Internet of Medical Things Data. Sensors, 21*(11), 3726. https://doi.org/10.3390/s21113726

Vaez, N., Desgualdo-Pereira, L., & Paglialonga, A. (2014). Development of a test of suprathreshold acuity in noise in Brazilian Portuguese: A new method for hearing screening and surveillance. *BioMed Research International, 2014,* 652838. https://doi.org/10.1155/2014/652838

Ventry, I. M., & Weinstein, B. E. (1983). Identification of elderly people with hearing problems. *ASHA, 25*(7), 37–42.

Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction, 3*(3), 615–661. https://doi.org/10.3390/make3030032

Watson, C. S., Kidd, G. R., Miller, J. D., Smits, C., & Humes, L. E. (2012). Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a U.S. version. *Journal of the American Academy of Audiology, 23*(10), 757–767. https://doi.org/10.3766/jaaa.23.10.2

World Health Organization. (1991). *Report of the informal working group on prevention of deafness and hearing impairment programme planning: Geneva, 18–21 June 1991*. https://apps.who.int/iris/handle/10665/58839

World Health Organization. (2021a). *Deafness and hearing loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

World Health Organization. (2021b). *World report on hearing*. https://www.who.int/publications/i/item/world-report-on-hearing

World Bank Open Data. (n.d.). *Population ages 65 and above (% of total population)*. https://databank.worldbank.org/reports.aspxsource=2&series=SP.POP.65UP.TO.SZ&country=

Zanet, M., Polo, E. M., Lenatti, M., van Waterschoot, T., Mongelli, M., Barbieri, R., & Paglialonga, A. (2021). Evaluation of a novel speech-in-noise test for hearing screening: Classification performance and transducers characteristics. *IEEE Journal of Biomedical and Health Informatics, 25*(12), 4300–4307. https://doi.org/10.1109/jbhi.2021.3100368

Zanet, M., Polo, E. M., Rocco, G., Paglialonga, A., & Barbieri, R. (2019). *Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing*. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), *2019,* 6991–6994. https://doi.org/10.1109/embc.2019.8857492