POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Biomedical Engineering

AUTOMATED DETECTION OF HEARING LOSS BY MACHINE LEARNING APPROACHES APPLIED TO SPEECH-IN-NOISE TESTING FOR ADULT HEARING SCREENING

Supervisors:

Prof. Alessia Paglialonga

Prof. Riccardo Barbieri

Co-supervisor:

Edoardo Maria Polo

Master thesis by:

Marta Lenatti, Matr: 916290

Academic year 2019-2020

Ringraziamenti

Vorrei ringraziare la professoressa Alessia Paglialonga e il professor Riccardo Barbieri per essere sempre stati disponibili in questi mesi ed avere reso possibile lo svolgimento di questo progetto nonostante lo stato di incertezza in cui fin dall'inizio ci siamo trovati.

Un grazie particolare a Edoardo e Marco per i preziosi consigli, i suggerimenti su come procedere e il sostegno 'virtuale'.

Ringrazio i miei genitori, che con i loro sacrifici hanno permesso il completamento di questo percorso universitario, sostenendomi in ogni decisione presa.

Alla mia famiglia e ai miei amici, grazie per aver sempre creduto in me.

Abstract

Hearing loss is the fourth leading cause of disability worldwide, with almost half a billion people affected. Despite its significant burden, hearing loss is often underestimated and it is rarely considered a handicap, especially among older adults that tend to seek assistance very late. The development of dedicated mobile apps may allow the quick and widespread diffusion of hearing screening, so to promote awareness and early assessment of the hearing conditions.

Pure-tone audiometry (PTA) is the gold standard in audiological testing. However, a series of limitations prevent it from being used on a large scale. In addition, audiometry is not able to evaluate the abilities of speech recognition in presence of noise, whose decrease is often associated with age-related hearing loss. In this context, speech-in-noise tests (SNTs) have been developed. This study aims at evaluating the classification performance of a newly developed SNT in the detection of mild or higher degree of hearing loss, according to the World Health Organization (WHO) criteria for hearing impairment. In particular, seven different supervised machine leaning approaches were trained on different sets of features extracted from the experimental procedure, including the score of the 'Hearing Handicap Inventory for the Elderly Screening Version' (HHIE-S) questionnaire, the age of the participant and other parameters directly related to the test execution. All the considered models achieved a moderate level of accuracy in the detection of hearing loss, with Support Vector Machines (SVMs) and logistic regression performing slightly better with respect to the others; comparable metrics were obtained with different sets of features, demonstrating that the exclusion of those features that were less informative, did not substantially change the overall classification performance. The results obtained match those of some of the most popular SNTs currently available, however further research is needed to validate the speech-in-noise test as an adult hearing screening test, on a much larger population involving a higher number of subjects with hearing loss, also with higher degrees of impairment.

Summary

Introduction: hearing loss, with almost half a billion people affected worldwide, is one of the major health concerns in ageing societies, causing communication problems, limiting social participation, and influencing the quality of life. Despite its high prevalence and strong social and emotional consequences, hearing loss is often an underrated disability, in fact hearing test are usually not included in routine healthcare screenings.

As a result, hearing loss in most cases remains undiagnosed, especially in older adults, as they perceive hearing loss as a natural consequence of ageing and, for this reason, they do not consider it as a handicap. Besides, recreational and occupational noise exposure is a major concern in the wide diffusion of hearing loss also among young individuals. However, most of the people wait until their hearing abilities deteriorate in a severe way before seeking for help, when an early diagnosis could have significantly limited the functional limitations generated by hearing loss. In order to promote awareness and early identification, hearing screening procedures should be widely accessible through the organization of appropriate campaigns (i.e. at school, in the workplace) or provided remotely.

Pure Tone Audiometry (PTA) is the gold standard for audiological clinical tests; however, it relies on the presence of trained healthcare operators, on the use of calibrated audiologic equipment, and it has to be performed in quiet environments. In addition, pure-tone hearing thresholds represent only a partial picture of hearing functionality as they do not capture the complex picture of hearing and communication, particularly speech communication. These limitations prevent PTA from being used on large-scale hearing screening initiatives.

Over the last few decades, a variety of speech-in-noise tests (SNTs) has been developed and validated as screening tools allowing to obtain a quick, costless, and self-administered hearing evaluation delivered via smartphone or online and therefore overcoming the limitations of audiometry.

A deterioration in speech recognition in noise is one of the consequences of both age-related hearing loss (presbycusis) and noise induced hearing loss (NIHL), which are among the most widespread types of hearing loss. All SNTs measure speech recognition in noise, however they differ a lot in terms of speech material (digits, CVC, VCV, words…), added background noise and testing procedure. Indeed, depending on the type of stimuli used, these tests may result more or less language-dependent.

Recently, a new automated speech-in-noise test based on a three-alternative forced-choice (3AFC) recognition task of meaningless Vowel-Consonant-Vowel (VCV) was developed from a collaboration between Politecnico di Milano and CNR, Consiglio Nazionale delle Ricerche. The test demonstrated high accuracy and test-retest repeatability, with no significant learning effects. Listening to the selected speech material (meaningless VCVs) does not involve substantial cognitive process and allows to obtain results only marginally influenced by the user's language and educational level, thus demonstrating that a potential mobile app based on this SNT may be used for widespread hearing screening.

The objective of the study was to assess the accuracy and the viability of the implemented procedure as an adult hearing screening tool, i.e. to promote a quick assessment of hearing conditions in people who would normally not seek any assistance.

Nowadays, machine learning approaches have been applied to a variety of applications, including audiology. In the present study, seven different supervised machine learning techniques were investigated, to compare their ability to detect hearing loss of mild or higher degree, as defined according to the World Health Organization (WHO) criteria based on the average PTA thresholds.

Materials and methods: 148 adults of different native languages and varying degrees of hearing loss were tested in non-clinical settings both in the laboratory and in several initiatives of local hearing screening. A subset of participants (8 out of 148) was tested in both ears whereas the remaining participants were tested in one ear, for a total of 156 ears tested.

The binary variable representing the screening results (pass/fail) was obtained based on the WHO criteria for hearing impairment, specifically based on the average pure-tone hearing threshold computed on the four central frequencies (0.5,1,2, and 4 kHz).

Due to the nature of hearing screening as a tool for early identification, mild-to-moderate degrees of hearing loss were considered in this study to address the ability of the test to identify earlier hearing difficulties. The target variable was set to 1 ('fail') if the WHO criterion for mild or moderate hearing impairment was satisfied (PTA> 25 dB HL for mild hearing impairment, referred to as criterion 1, and PTA> 40 dB HL for moderate hearing impairment, referred to as criterion 2) or 0 ('pass') otherwise.

The dataset used for classification includes 156 records, corresponding to the different ears tested, and 11 columns, including the following features: the speech reception threshold ('SRT'), the raw score of the Hearing Handicap Inventory for the Elderly Screening Version (HHIE-S) questionnaire ('Score'), the age ('Age'), the number of trials ('#Trials'), the number of correct responses ('#Correct'), the percentage of correct responses achieved ('%Correct'), the average reaction time ('Avg_reaction_time'), the total test duration ('Total_test_time'), the volume level ('Volume'), the classification into three classes of hearing impairment according to the HHIE-S questionnaire raw score ('Score_classes') and the target variable previously introduced.

These variables represent a set of features extracted from the information gathered during the three phases of the experimental procedure: PTA assessment, personal information and HHIE-S questionnaire compilation and test execution. The SRT, defined as the signal-to-noise ratio (SNR) at a certain intelligibility level, gives a clue about the ability to recognize speech in noise, thus it is the usual main outcome of speech-in-noise tests. However, the addition of other features to implement a multivariate classification model may lead to better performances in discriminating the presence of hearing impairment.

The Spearman's correlation coefficients were analyzed to evaluate the strength and the direction of the relationship between couples of attributes and also between attributes and target variable, in order to obtain a first hint about the presence of multicollinearity.

Then, scatterplots were analyzed in the pass/fail classes to visualize which couples of variables were more able to discriminate between the two classes and which ones were not.

The performance of SRT as predictor of mild and moderate hearing loss was assessed by analyzing ROC curves to evaluate the discrimination capabilities as a function of the cut-off SRT value.

Next, generalized linear models (GLMs) were implemented for each single attribute, in order to evaluate the contributions of the various features in discriminating the screening result and to identify the most meaningful variables and those that might be disregarded in a classification task. Afterwards, GLMs were addressed by adding SRT as a second predictor and the resulting interaction term.

As a last step, seven different classifiers, including four of the most widespread and consolidated approaches (Decision Trees (DT), Support Vector Machines (SVM), logistic regression, K-Nearest-Neighbors) and three ensemble methods (ensemble logistic

regression, Random forests and Gradient boosting) were compared in terms of performance in detecting hearing loss based on the WHO criterion for mild hearing impairment (criterion 1). Their classification performance was evaluated by means of a variety of metrics including the proportion of records correctly classified as 'pass' or 'fail' (training and test accuracy), the proportion of subjects with no hearing loss ('pass') correctly classified (specificity) and the proportion of subjects with hearing impairment ('fail') correctly classified (sensitivity).

Results: analysis of the possible influence of sex, ear tested (left/right), and age on the on the collected data revealed no significant effect of sex and ear and a significant effect of age, with higher age associated in worse test outcomes.

Three age-related groups were considered, namely 'Young' (i.e., age: 20 - 25 years), 'Adults' (i.e., age: 25 - 60 years), and 'Elderly' (i.e., age $\geq$ 60 years).

As a result, speech recognition abilities in noise worsen, as reported by SRT that from a mean value of -15.86 dB SNR in the Young group increases to -12.86 dB SNR in the adults and finally increases again to -6 dB SNR in elderly. In addition, also the average pure-tone hearing threshold passes from an average value of 0 dB HL in young subjects to a value of about 20 dB HL in adults and finally to a much higher value, of about 31 dB HL, in elders.

The duration of the test, instead, seems to remain pretty constant regardless of age as younger users require an higher number of stimuli before finishing the test, but they respond faster to each VCV proposal, whereas older subjects may respond to fewer stimuli, as they are likely to perform worse in the test and thus reach the stop criterion (i.e. 12 reversals) earlier, but each time they will spend more time selecting the single response.

Regarding multicollinearity among classification attributes, the chosen test volume showed no significant correlation with the other variables, except for a weak negative correlation with the average reaction time. SRT instead showed a moderate correlation with almost all the other variables, and a strong correlation (r=0.83) was observed only with the percentage of correct responses.

The correlation between SRT and the binary result of the PTA is moderate (r=0.66) and in line with the values reported in the literature for similar online SNTs such as Earcheck and Occupational Earcheck.

GLMs evaluating SRT against the binary PTA showed a significant contribution of SRT for both mild and moderate hearing impairment detection. Indeed, SRT is one of the best

candidates for discriminating the presence of hearing loss as defined from pure-tone hearing thresholds.

The variable most strongly correlated with the target is the age (r=0.76), however also other features such as the number of correct answers and the average reaction time are quite correlated with the screening result (r= -0.62 and 0.53, respectively), suggesting that they also may be useful attributes for classification purposes.

The scatterplots of paired variables have been evaluated against the capability to separate 'pass' and 'fail' records. Features like the SRT, age, the number of correct answers and the average response time allow to have a fairly well defined separation between the two classes for both criterion 1 and criterion 2, as pass instances are gathered in correspondence of lower (i.e. better) SRTs, older age, higher number of correct responses, lower single response reaction time and so on.

Conversely, other parameters such as test total execution time and volume have shown less ability to separate the two categories, since the distribution of records in the two classes is much wider and without obvious relationship with the couple of variables considered.

ROC curves were built progressively varying the discrimination SRT and by selecting the cut-off SRT as the value closest to the ideal point (i.e. top left angle of the curve). As far as it concerns criterion 1, an SRT discrimination threshold of -8.875 dB SNR yielded an AUC equal to 0.83, a specificity of 0.84, a sensitivity of 0.76, and a test accuracy of 0.81. Considering criterion 2, a cut-off SRT of -6 dB SNR yielded better values in terms of AUC (0.89), sensitivity (0.89), and accuracy (0.83), but slightly lower specificity (0.81).

GLMs evaluating the other attributes against the binary PTA showed that, besides the volume and the total test duration, all the other variables contributed significantly to the prediction for the two WHO criteria for hearing impairment here considered. However, the amount of variance of the dependent variable explained by the single predictor was quite low, as a maximum value of 0.40 was observed (predictor: age; criterion: 1). An improved goodness-of-fit of the GLM was achieved by using models that included the SRT and each one of the other features as predictors, reaching the maximum value of adjusted R^2 equal to 0.44 (predictors: age, SRT; criterion: 1). The interactions between the attributes and SRT were in general not significant, and even when they were significant, they did not bring tangible improvements to the model, suggesting limited or no influence of the interaction terms on the predicted variable.

Based on the results of features characterization and GLMs described above, 7 different sets of features were considered and fed to 7 different supervised machine leaning approaches (Decision Trees (DT), Support Vector Machines (SVM), logistic regression, K-Nearest-Neighbors, ensemble logistic regression, Random forests and Gradient boosting).

In this preliminary evaluation, logistic regression and Support Vector Machines obtained the best results in terms of classification performance, with an AUC of 0.91, an accuracy on the test set equal to 0.82, a specificity of 0.82, and a sensitivity of about 0.8.

Random forest and Gradient boosting techniques were promising classifiers too as they showed the same specificity as the previously mentioned methods (i.e., 0.82) and slightly lower AUC (0.87) and accuracy on the test set (0.79 and 0.78, respectively). However, their ability to detect hearing impaired subjects, represented by sensitivity, appears to be too low (0.73 and 0.7).

KNN presented an AUC equal to 0.85 and a specificity of 0.77 which resulted to be lower than that of the previous methods. However, it yielded a test accuracy comparable to the one found for Gradient boosting (i.e., 0.77) and a better sensitivity (0.77) with respect to Random Forest and Gradient boosting.

Unfortunately, the ensemble logistic regressor underwent overfitting, as can be seen from the big difference between training (0.95) and test (0.78) accuracy, and for this reason it was not feasible for this preliminary assessment because it may be prone to significant prediction errors on new instances.

Regarding DTs, they have the advantages of being the fastest approach and giving more interpretable results. However, the classification performance of DTs in this preliminary study seemed to be the lowest, especially in terms of AUC (0.74). The AUC obtained with DTs using a reduced set of 4 features is higher (0.82), but still lower compared to other classification algorithms.

Discussion and conclusions: significant worsening of the test parameters has been observed with increasing age considering the three groups (Young, Adults and Elderly). Indeed, young subjects (i.e., age: 20 - 25 years) showed excellent speech recognition abilities in noise; in the Adults group (i.e., age: 25 - 60 years) the performance was moderate, and in considering Elderly (i.e., age $\geq$ 60 years) the performance had a consistent worsening.

In fact, features like SRT, the average pure-tone thresholds, the HHIE-S questionnaire score and the average reaction time tend to increase whereas the number of trials, the number of correct responses and consequently the percentage of correct responses tend to diminish as age increases, showing an overall decrease in the test outcomes and therefore poorer capabilities to recognize speech in a background noise. Indeed, increased hearing thresholds and decreased speech recognition ability, particularly in noise, are among the first symptoms of age-related hearing impairment.

The discrimination capabilities for criterion 1, considering classification of ears based on the measured SRT, were fully comparable to the previous findings on a smaller population (i.e., 98 subjects), using cut-off values of -8 dB SNR and -10 dB SNR.

Comparing the observed performance for criterion 1 and criterion 2, the latter showed better values in terms of AUC, sensitivity, and accuracy and slightly lower specificity. The classification performance for criterion 2 matches the measures obtained for the Speech Understanding in Noise (SUN) test, another SNT based on the same speech material and recognition task (three-alternatives forced choice) and a fixed-levels (i.e., non-adaptive) testing procedure. However, it should be noted that the number of records belonging to the 'fail' class for criterion 2 includes only a few subjects in the tested population (i.e., 18), therefore these outcomes need further demonstration after a higher number of subject with moderate hearing impairment is tested in future studies.

Regarding the analysis of the full set of variables, evaluations of the correlation matrix and of matrices of scatterplots of paired features suggested that features like volume and test duration might not bring important contributions to classification of hearing impairment. These considerations were further confirmed by the analysis of single-predictor GLMs on these features and multiple-predictor GLMs (i.e., including SRT and each of these features). The volume may be disregarded from the set of features in this preliminary assessment as it is less informative, perhaps also related to the fact that few participants decided to adjust the volume, whereas the majority maintained the default value as they considered it suitable for test execution.

The test duration also seems to be a negligible feature for classification purposes as it does not substantially change with varying hearing thresholds in the tested population. This is related to a compensatory action as people who perform well in the test have to go through more trials (i.e., to reach a lower SRT) before reaching the end of the test, but they tend to

respond more quickly at each trial because they are less hesitant in VCV recognition. On the other hand, individuals with worse speech recognition performance have to go through fewer trials, but they could take on average the same total time as good performers, because of their longer response time due to a higher hesitation in giving their answers. As a result, subjects tend to take more or less the same time to perform the test, almost independently on test outcomes.

On the other hand, due to the above considerations, the average time needed to reply to the proposed stimulus may be a more interesting parameter to discriminate the screening result, as it may indirectly reflect the subject's ability to recognize the stimuli. In addition to the features previously considered, the removal of the HHIE-S questionnaire outcomes can accelerate the experimental procedure and reduce the presence of possible bias that is frequently observed in questionnaires and might also be related to a reduced self-perception of the presence of hearing problems, mostly observed in the elderly.

Comparable performance was obtained with different sets of features , demonstrating that the exclusion of those features that were less informative, or those that showed high correlation with other features, did not substantially change the overall classification performance, therefore exclusion of unnecessary features may be helpful, for example to speed up the training of the algorithm and to reduce overfitting. Among the different methods here considered, SVM and logistic regression seemed to be the more promising approaches. They performed better in terms of sensitivity and AUC as compared to the classification based on SRT only.

The performance observed with Support Vector Machine and logistic regression algorithms reached a moderate level of accuracy, in line with the findings obtained for other well-known automated speech-in-noise tests. However, the sensitivity needs to be further improved because about one fifth of the examined subjects with hearing loss are erroneously classified as normal hearing.

Moreover, the selection of training and test partitions in small datasets like the one here considered, may introduce variability in the classification. In this study, to address the potential variance due to changes in the underlying data, the average performance across 1000 iterations of the model was considered. Results demonstrated that the average performance of the models was similar, both considering different classification algorithms and different sets of features; the standard deviations were relatively low (i.e., smaller than

0.1) for parameters like the accuracies and AUC, reached values around 0.1 for F-measure and specificity and slightly higher values for FNR, precision, and sensitivity (e.g. 0.17, 0.12 and 0.17 considering gradient boosting). Because of this intrinsic variability, the performance of a single optimized model may differ consistently from the average classification performance. Additional studies on larger samples may therefore allow to minimize this variability, by reducing the data-dependency, as each partition tends to be more equally representative of the original dataset when the sample size is substantially increased. Subsequently, optimized model definition and finer hyperparameter tuning could be achieved and this could further help towards the development of a smartphone application for adult hearing screening.

In general, further research is needed to validate the speech-in-noise test and the classification methods for both criterion 1 and criterion 2, on a much larger population involving a higher number of subjects with hearing impairment, also with higher degrees of hearing loss.

Keywords: hearing loss; hearing screening; pure-tone audiometry; speech-in-noise test; speech reception threshold; supervised learning; classification; machine learning; generalized linear models

Abstract

La perdita dell'udito è la quarta causa di disabilità al mondo, con quasi mezzo miliardo di persone colpite; tuttavia, essa è spesso sottovalutata e raramente viene considerata un handicap, soprattutto tra gli anziani che tendono a cercare assistenza troppo tardi. Lo sviluppo di app dedicate può consentire una vasta diffusione degli screening audiologici, in modo da promuovere una rapida valutazione delle condizioni dell'udito.

L'audiometria tonale è il gold standard nei test audiologici, tuttavia, una serie di limitazioni ne impedisce l'uso su larga scala. Inoltre, essa non è in grado di valutare le capacità di riconoscimento del parlato nel rumore, il cui peggioramento è spesso associato alla perdita dell'udito legata all'età. In questo contesto, sono stati sviluppati diversi test di speech-in-noise (SNT). Questo studio mira a valutare la performance di classificazione di un SNT recentemente sviluppato, nella rilevazione di perdite uditive, definite secondo i criteri dell'Organizzazione Mondiale della Sanità (OMS). Sette diversi approcci di apprendimento supervisionato sono stati addestrati utilizzando una serie di caratteristiche estratte dalla procedura sperimentale, tra cui il punteggio del questionario "Hearing Handicap Inventory for the Elderly Screening Version" (HHIE-S), l'età del partecipante e altri parametri collegati all'esecuzione del test. Tutti i modelli considerati hanno raggiunto un buon livello di accuratezza nella rilevazione delle perdite dell'udito, con risultati migliori associati a Support Vector Machines (SVM) e regressione logistica. Utilizzando diversi set di caratteristiche sono state ottenute metriche comparabili, a dimostrazione del fatto che l'esclusione delle variabili meno informative non modifica la performance complessiva della classificazione. I risultati ottenuti corrispondono a quelli di alcuni tra i più popolari SNT attualmente disponibili, tuttavia ulteriori ricerche sono necessarie per convalidare il test come strumento di screening dell'udito per adulti, su una popolazione molto più ampia che coinvolga un numero maggiore di soggetti con perdite uditive, anche con gradi di compromissione più elevati.

Sommario

Introduzione: la perdita dell'udito, con quasi mezzo miliardo di persone colpite in tutto il mondo, è una delle maggiori preoccupazioni sanitarie nelle società in cui l'aspettativa di vita sta aumentando, ed è causa di problemi di comunicazione, limiti alla partecipazione sociale e riduzione della qualità della vita. Nonostante la sua elevata diffusione e le forti conseguenze sociali ed emotive, la perdita dell'udito è spesso una disabilità sottovalutata, infatti i test dell'udito non sono solitamente inclusi nelle visite mediche di routine.

Di conseguenza, nella maggior parte dei casi, la perdita dell'udito non viene diagnosticata, soprattutto negli adulti più anziani, in quanto essi la percepiscono come una conseguenza naturale dell'invecchiamento e per questo motivo non la considerano un handicap.

Inoltre, l'esposizione al rumore a scopo ricreativo e professionale è una delle principali cause nell'ampia diffusione della perdita dell'udito anche tra i giovani. Tuttavia, la maggior parte delle persone attende che le loro capacità uditive si deteriorino in modo grave prima di cercare aiuto, quando una diagnosi precoce avrebbe potuto limitare in modo significativo la limitazione funzionale causata dalla perdita uditiva. Al fine di promuovere la sensibilizzazione e l'identificazione precoce, le procedure di screening dell'udito dovrebbero essere ampiamente accessibili attraverso l'organizzazione di apposite campagne (ad es. a scuola, sul posto di lavoro…) o proposte in modalità da remoto.

L'Audiometria Tonale (PTA) è considerata il test clinico audiologico standard; tuttavia, essa si basa sulla presenza di operatori sanitari preparati, sull'uso di apparecchiature audiologiche calibrate e necessita di essere eseguita in ambienti silenziosi. Inoltre, le soglie uditive dei toni puri non sono in grado di rappresentare le funzionalità uditive nella loro totalità, poichè non colgono il complesso scenario dell'udito e della comunicazione, in particolare la comunicazione verbale. Queste limitazioni impediscono l'utilizzo della PTA in iniziative di screening uditivo su larga scala. Negli ultimi decenni, una serie di test di speech-in-noise (SNT) ("parlato nel rumore") sono stati sviluppati e validati come strumenti di screening in grado di ottenere una valutazione dell'udito rapida, economica e autosomministrata tramite smartphone oppure online, superando i limiti dell'audiometria.

Il deterioramento del riconoscimento del parlato nel rumore è una delle conseguenze sia della perdita dell'udito dovuta all'età (presbiacusia) che della perdita dell'udito indotta da una prolungata esposizione al rumore, tra i tipi di perdita uditiva più diffusi.

Tutti gli SNT misurano il riconoscimento vocale nel rumore, tuttavia differiscono molto in termini di materiale vocale (cifre, CVC, VCV, parole...), rumore di fondo aggiunto e procedura di test. Per questo, a seconda degli stimoli utilizzati, questi test possono risultare più o meno dipendenti dalla lingua.

Recentemente, un nuovo test automatizzato di speech-in-noise basato sul riconoscimento, a partire da tre alternative (3AFC), di sequenze vocale-consonante-vocale (VCV) prive di significato, è stato sviluppato da una collaborazione tra il Politecnico di Milano e il Consiglio Nazionale delle Ricerche (CNR). Il test ha dimostrato un'elevata accuratezza e ripetibilità, con effetti di apprendimento non significativi. L'ascolto del materiale vocale selezionato (VCV privi di significato) non comporta alcun processo cognitivo concreto e permette di ottenere risultati influenzati solo marginalmente dal linguaggio dell'utente e dal suo livello di istruzione, dimostrando che un'applicazione mobile basata su questo SNT potrebbe essere distribuita e utilizzata per screening uditivi diffusi su larga scala.

L'obiettivo dello studio è quello di valutare l'accuratezza e la fattibilità della procedura implementata come test di screening dell'udito per adulti, per promuovere una rapida valutazione delle condizioni di udito in persone che normalmente non richiederebbero alcuna assistenza.

Al giorno d'oggi, gli approcci di machine learning sono stati applicati a una varietà di discipline, inclusa l'audiologia. Nel presente studio sono state considerate sette diverse tecniche di machine learning, allo scopo di confrontare la loro abilità nel rilevare perdite uditive lievi oppure più severe, definite secondo i criteri dell'Organizzazione Mondiale della Sanità (OMS) sulla base delle soglie uditive medie.

Materiali e metodi: un gruppo di 148 adulti di diversa lingua nativa e con diverso grado di perdita dell'udito è stato testato in contesti non clinici, sia in laboratorio che in diverse iniziative di screening uditivo a livello locale. Alcuni partecipanti (8 su 148) sono stati testati su entrambi le orecchie, mentre i rimanenti sono stati testati su un orecchio solo, per un totale di 156 orecchie testate.

La variabile binaria, rappresentante il risultato dello screening (pass / fail), è stata ottenuta a partire dai criteri dell'OMS per l'ipoacusia, definiti utilizzando la soglia media uditiva dei toni puri sulle quattro frequenze centrali (0.5, 1, 2, e 4 kHz). A causa della natura dello screening dell'udito come strumento per l'identificazione precoce, solo perdite uditive meno

gravi (da lievi a moderate) sono state considerate in questo studio, per valutare l'abilità del test di riscontrare delle perdite uditive ai primi stadi.
La variabile target è stata impostata a 1 ("fail") nel caso in cui il criterio dell'OMS per perdite di udito lievi o moderate fosse soddisfatto (PTA> 25 dB HL per perdite uditive lievi, denominate criterio 1 e PTA> 40 dB HL per perdite uditive moderate, definite come criterio 2) o 0 ("pass") altrimenti.

Il set di dati raccolto è costituito da 156 righe, corrispondenti alle diverse osservazioni, e 11 colonne, che includono le seguenti variabili: la speech reception threshold ('SRT'), il punteggio del questionario 'Hearing Handicap Inventory for the Elderly Screening Version' (HHIE-S) ('Score'), l'età ('Age'), il numero di trials ('#Trials'), il numero di risposte corrette ('#Correct'), la percentuale di risposte corrette ('%Correct'), il tempo medio di reazione ('Avg_reaction_time'), la durata totale del test ('Total_test_time'), il livello del volume ('Volume'), la suddivisione in tre gradi di perdita uditiva a seconda del punteggio ottenuto nel questionario HHIE-S ('Score_classes') ed infine la variabile target precedentemente introdotta.
Queste variabili rappresentano un insieme di caratteristiche estratte dalle informazioni raccolte durante le tre fasi della procedura sperimentale: valutazione audiometrica, compilazione di informazioni personali e del questionario HHIE-S ed esecuzione del test.

L'SRT, definito come il rapporto segnale-rumore ad un certo livello di intelligibilità, offre un indizio sulla capacità di riconoscere il parlato nel rumore, quindi è il risultato principale comunemente fornito nei test di speech-in-noise. Tuttavia, l'aggiunta di altri attributi per implementare un modello di classificazione multivariata può portare a migliori prestazioni nella discriminazione della presenza di deficit uditivo.

I coefficienti di correlazione di Spearman sono stati analizzati per valutare la presenza e l'entità delle relazioni tra le coppie di attributi e anche tra attributi e variabile target, al fine di ottenere un primo indizio sulla presenza di multicollinearità.
Successivamente, gli scatterplot in funzione delle classi "pass" e "fail" sono stati analizzati per visualizzare quali coppie di variabili fossero maggiormente in grado di discriminare tra le due classi e quali no.

La performance dell'SRT come predittore di perdite uditive lievi e moderate è stata in seguito valutata analizzando delle curve ROC allo scopo di valutare le capacità di discriminazione in funzione del valore della SRT.
Modelli Lineari Generalizzati (GLM) sono stati implementati per ogni singolo attributo al fine di valutare il contributo delle varie caratteristiche nella discriminazione del risultato dello screening e identificare quali fossero le variabili più significative e quali variabili potessero essere ignorate in un task di classificazione. Successivamente, ulteriori GLM sono stati analizzati, aggiungendo SRT come secondo predittore ed infine anche il relativo termine di interazione.
Come ultimo passo, sette diversi classificatori, tra cui quattro dei più diffusi e consolidati approcci (alberi decisionali, Support Vector Machines, regressione logistica, K-Nearest-Neighbors) e tre metodi ensemble (regressione logistica ensemble, Random forests e Gradient Boosting) sono stati confrontati in termini di performance nell'identificare le perdite uditive sulla base del criterio dell'OMS per i problemi di udito lievi (criterio 1).
La loro prestazione nella classificazione è stata valutata attraverso una serie di metriche, tra cui la percentuale di record correttamente classificati, sia "pass" che "fail", nelle partizioni del dataset (accuratezza dell'addestramento e del test), la quota di soggetti senza ipoacusia ("pass") correttamente classificati (specificità) e la quota di soggetti con ipoacusia ("fail") correttamente classificati (sensibilità).

Risultati: l'analisi della possibile influenza del sesso, dell'orecchio testato (destro/sinistro) e dell'età sui dati raccolti ha rivelato che nessun effetto significativo è presente per quanto riguarda il sesso dell'utente né l'orecchio testato, mentre un effetto significativo è presente per quanto riguarda l'età, in quanto l'aumento dell'età è associato ad un peggioramento nei risultati del test.

Sono stati considerati tre gruppi: 'Giovani' (i.e., età: 20 - 25 anni), 'Adulti' (i.e., età: 25 - 60 anni) e 'Anziani' (i.e., età $\geq$ 60 anni).
Le capacità di riconoscimento vocale nel rumore peggiorano, come riportato dall'SRT, che da un valore medio di -15,86 dB SNR nel gruppo dei giovani aumenta a -12,86 dB SNR negli adulti e infine aumenta di nuovo a -6 dB SNR negli anziani. Inoltre, anche la soglia uditiva media dei toni puri passa da un valore medio di 0 dB HL nei soggetti giovani ad un

valore di circa 20 dB HL negli adulti ed infine ad un valore molto più alto, di circa 31 dB HL, negli anziani.
La durata del test, invece, sembra rimanere piuttosto costante, indipendentemente dall'età, in quanto gli utenti più giovani richiedono un numero maggiore di stimoli prima di finire il test, ma rispondono più velocemente ad ogni proposta di VCV, mentre i soggetti più anziani in genere rispondono a meno stimoli in quanto forniscono in media prestazioni peggiori nel test e quindi raggiungono il criterio di stop (cioè 12 inversioni) prima, ma ogni volta impiegheranno più tempo a selezionare la singola risposta.
Per quanto riguarda la multicollinearità tra gli attributi di classificazione, il volume di test selezionato non ha mostrato alcuna correlazione significativa con le altre variabili, ad eccezione di una debole correlazione negativa con il tempo di reazione medio.
L'SRT ha mostrato invece una correlazione moderata con quasi tutte le altre variabili, e una forte correlazione (r=0.83) è stata osservata solo con la percentuale di risposte corrette.
La correlazione tra SRT e il risultato binario della PTA è moderata (r=0.66) e in linea con i valori riportati in letteratura per altri SNT online simili, quali Earcheck e Occupational Earcheck.
I modelli lineari generalizzati implementati per valutare l'SRT rispetto alla PTA binaria hanno mostrato un contributo significativo dell'SRT sia per la rilevazione di deficit uditivo lieve che moderato. Per questo, l'SRT è uno dei migliori candidati per discriminare la presenza di ipoacusia come definita dall'OMS a partire dalle soglie uditive dei toni puri.
La variabile più fortemente correlata con il target è l'età (r=0.76), tuttavia anche altri attributi, come il numero di risposte corrette e il tempo medio di reazione, sono abbastanza correlate con il risultato dello screening (r= -0.62 e 0.53, rispettivamente), suggerendo che anch'esse possano essere variabili utili ai fini della classificazione.

Gli scatterplot di variabili accoppiate sono stati valutati rispetto alla capacità di separare i record "pass" e "fail". Caratteristiche come l'SRT, l'età, il numero di risposte corrette e il tempo di risposta medio consentono di avere una separazione abbastanza definita tra le due classi, sia per il criterio 1 che per il criterio 2, in quanto le istanze "pass", ad esempio, sono raccolte in corrispondenza di valori di SRT bassi (cioè migliori), età più avanzata, numero più alto di risposte corrette, tempo di reazione di una singola risposta più basso...
Al contrario, altri parametri quali il tempo di esecuzione totale del test e il volume hanno mostrato una minore capacità di separare le due categorie, poiché la distribuzione delle

osservazioni nelle due classi è molto più ampia e senza evidenti relazioni con la coppia di variabili considerate.

Le curve ROC sono state costruite variando progressivamente la SRT e selezionando la SRT soglia come il valore più vicino al classificatore ideale (cioè l'angolo in alto a sinistra della curva ROC). Per quanto riguarda il criterio 1, una $SRT_{cut\text{-}off}$ di -8.75 dB SNR ha prodotto un AUC pari a 0.83, una specificità di 0.84, una sensibilità del 0.76 e una accuratezza di 0.81. Considerando il criterio 2, una $SRT_{cut\text{-}off}$ di -6 dB SNR ha prodotto valori migliori in termini di AUC (0.89), sensibilità (0.89) e accuratezza (0.83), ma una specificità leggermente inferiore (0.81).

I GLM che valutano gli altri attributi rispetto alla PTA binaria hanno mostrato che, ad eccezione del volume e della durata totale del test, tutte le altre variabili contribuiscono significativamente alla predizione dei danni uditivi, per quanto riguarda i due criteri OMS per i deficit uditivi qui considerati. Tuttavia, l'entità della varianza della variabile dipendente spiegata dal singolo predittore è risultata essere piuttosto bassa, con un valore massimo osservato pari a 0.40 (predittore: età; criterio: 1). Un miglioramento della bontà del GLM è stato ottenuto utilizzando modelli che includevano la SRT e ciascuna delle altre variabili come predittori, raggiungendo un valore massimo di R^2 aggiustato, pari a 0.44 (predittori: età, SRT; criterio: 1).

Le interazioni tra gli attributi e SRT non sono risultate in generale significative e anche nei casi in cui lo erano, non hanno portano miglioramenti tangibili al modello, suggerendo una influenza limitata, oppure nulla, dei termini di interazione nei confronti della variabile predetta. Sulla base dei risultati della caratterizzazione degli attributi e dei GLM precedentemente descritti, 7 diversi set di attributi e 7 diversi approcci di apprendimento supervisionato (Alberi decisionali (DT), Support Vector Machines (SVM), regressione logistica, K-Nearest-Neighbor, regressione logistica ensemble, Random forests and Gradient boosting) sono stati considerati.

In questa valutazione preliminare, la regressione logistica e le Support Vector Machines sembrano ottenere i migliori risultati in termini di prestazioni di classificazione, con un AUC di 0.91, una accuratezza sul test set pari a 0.82, una specificità di 0.82 e una sensibilità intorno a 0.8.

Anche le tecniche Random forest e Gradient boosting hanno dimostrato di essere classificatori promettenti, avendo ottenuto la stessa specificità dei metodi precedenti (i.e.,

0.82) e valori leggermente minori in termini di AUC (0.87) e accuratezza sulla partizione di test (rispettivamente di 0.79 e 0.78). Tuttavia, la loro capacità di rilevare i soggetti con perdite uditive, rappresentata dalla sensibilità, è risultata essere troppo bassa (0.73 e 0.7).
KNN ha prodotto un AUC pari a 0.85 e una specificità di 0.77, che risultano essere inferiori rispetto alle tecniche precedenti. Tuttavia, ha fornito una accuratezza di test pari a quella riscontrata per il Gradient boosting (i.e., 0.77) e una migliore sensibilità (0.77) rispetto a Random forest e a Gradient boosting.
Purtroppo, il metodo di regressione logistica ensemble è andato incontro ad overfitting, come si può notare dalla grande differenza tra l'accuratezza di training (0.95) e quella di test (0.78) e per questo motivo non è risultato essere praticabile in questa valutazione preliminare, poiché potrebbe portare a errori di predizione significativi sui nuovi record.
Per quanto riguarda gli alberi decisionali, essi possiedono il vantaggio di essere l'approccio più veloce e di fornire i risultati maggiormente interpretabili. Tuttavia, la performance di classificazione dei DT, in questo studio preliminare, è risultata essere la peggiore, soprattutto in termini di AUC (0.74). L'AUC ottenuta con alberi decisionali basati su un set di sole 4 caratteristiche è più alta (0.82), ma ancora minore rispetto a quella ottenuta per altri algoritmi di classificazione.

Discussione e conclusioni: è stato osservato un significativo peggioramento dei parametri del test all'aumentare dell'età, considerando tre gruppi (Giovani, Adulti e Anziani).
I soggetti giovani (i.e., età: 20-25 anni) hanno mostrato eccellenti capacità di riconoscimento del parlato nel rumore; nel gruppo Adulti (i.e., età: 25-60 anni) le prestazioni risultano moderate e nel considerare gli Anziani (i.e., età $\geq$ 60 anni) le prestazioni hanno un consistente peggioramento. Infatti, caratteristiche come l'SRT, le soglie medie dei toni puri, il punteggio del questionario HHIE-S e il tempo di reazione medio tendono ad aumentare mentre il numero di prove, il numero di risposte corrette e di conseguenza la percentuale di risposte corrette tende a diminuire all'aumentare dell'età, mostrando un peggioramento generale nei risultati del test e di conseguenza capacità più povere nel riconoscere il parlato nel rumore. Invero, l'aumento delle soglie uditive e una diminuzione nelle capacità di riconoscimento verbale, specialmente nel rumore, sono tra i primi sintomi dei deficit uditivi legati all'età.
Le capacità di discriminazione ottenute per il criterio 1, considerando la classificazione degli individui sulla base della SRT misurata, sono pienamente paragonabili a risultati ottenuti in

precedenza considerando un campione ridotto di soggetti (i.e., 98 soggetti) e ottenendo valori di cut-off simili (i.e., -8 dB SNR e -10dB SNR).

Confrontando le performance osservate per il criterio 1 e il criterio 2, per quest'ultimo sono stati riscontrati valori migliori in termini di AUC, sensibilità e accuratezza e valori leggermente inferiori di specificità. Le prestazioni di classificazione dell'SRT per il criterio 2 corrispondono alle misure ottenute per il test Speech Understanding in Noise, un altro SNT basato sullo stesso materiale vocale e lo stesso task di riconoscimento (scelta forzata a tre alternative) e una procedura di test a livelli fissi, ovvero non adattativa. Tuttavia, è necessario notare che il numero di record della popolazione testata, appartenenti alla classe "fail" per il criterio 2 comprende solamente pochi soggetti (i.e., 18), pertanto questi risultati necessitano di essere ulteriormente dimostrati dopo che un numero più elevato di soggetti con problemi di udito moderato sarà testato in studi futuri.

Per quanto concerne l'analisi dell'intero set di variabili, la valutazione della matrice di correlazione e delle matrici di scatterplot ha suggerito che il volume e la durata totale del test non portano alcun contributo importante nella classificazione delle ipoacusie.

Queste considerazioni sono state ulteriormente confermate dall'analisi dei GLM a singolo predittore, su queste features e dai GLM a più predittori (i.e., che includevano la SRT e ciascuna di queste features).

In queste analisi preliminari il volume può essere escluso dall'insieme degli attributi, in quanto risulta essere meno informativo, probabilmente poiché solo alcuni partecipanti hanno deciso di aggiustarne il valore in base alle proprie esigenze, mentre la maggioranza ha mantenuto il valore di default in quanto lo ha ritenuto adatto per l'esecuzione del test.

Anche la durata totale del test sembra essere una feature trascurabile ai fini della classificazione, poiché non varia in maniera sostanziale al variare delle soglie uditive, nella popolazione testata. Ciò è legato ad un'azione di compensazione, dovuta al fatto che le persone che ottengono buoni risultati nel test devono compiere un numero maggiore di prove prima di giungere alla fine del test, ma rispondono in modo più rapido a ciascuna prova, in quanto sono meno esitanti nell'identificare i VCV. D'altra parte, gli individui con prestazioni di riconoscimento verbale peggiori, eseguono meno prove, ma possono impiegare in media lo stesso tempo totale di test di coloro che hanno prestazioni migliori, poiché hanno tempi di risposta maggiori per via della loro maggiore esitazione nel dare le risposte. Di

conseguenza, i soggetti tendono ad impiegare più o meno lo stesso tempo per eseguire il test, quasi indipendentemente dai risultati ottenuti nel test.
Al contrario, per via delle considerazioni fatte qui sopra, la selezione del tempo medio necessario per rispondere ad uno stimolo può essere un parametro interessante per discriminare il risultato dello screening, in quanto può indirettamente riflettere la capacità del soggetto di riconoscere gli stimoli.
Oltre alle caratteristiche precedentemente considerate, la rimozione del risultato del questionario HHIE-S può accelerare la procedura sperimentale e ridurre la presenza di possibili distorsioni che sono frequentemente osservate nei questionari e potrebbero essere dovute a una ridotta auto-percezione della presenza di problemi uditivi, maggiormente osservata negli anziani.
Prestazioni comparabili sono state ottenute con diversi set di attributi, dimostrando che l'esclusione delle caratteristiche meno informative o di quelle che mostravano una maggiore correlazione con altre caratteristiche, non ha influenzato le prestazioni complessive della classificazione, ma può al contrario essere utile, ad esempio, per accelerare l'addestramento dell'algoritmo e per ridurre l'overfitting.
Tra i diversi approcci qui considerati, SVM e la regressione logistica sembravano essere gli approcci più promettenti. Essi hanno inoltre ottenuto risultati migliori in termini di sensibilità e AUC rispetto alla classificazione basata sulla sola SRT.
La performance osservata con algoritmi di Support Vector Machine e regressione logistica ha raggiunto un livello di accuratezza moderato, in linea con i risultati ottenuti per altri tra i più diffusi test di speech-in-noise automatici. Tuttavia, la sensibilità necessita di essere ulteriormente migliorata, poiché circa un quinto dei soggetti esaminati che presenta una perdita uditiva è erroneamente classificato come normo udente.

Inoltre, la selezione delle partizioni di training e di test in dataset di piccole dimensioni, come quello qui considerato, può introdurre delle variabilità nella classificazione. In questo studio, la potenziale varianza dovuta a cambiamenti nei dati è stata trattata considerando la performance media su 1000 iterazioni del modello. I risultati hanno evidenziato performance medie simili sia considerando diversi algoritmi di classificazione che diversi set di attributi. Le deviazioni standard ottenute sono relativamente basse (minori di 0.1) per parametri come accuratezza e AUC, hanno raggiunto valori intorno a 0.1 per la F-measure e la specificità e valori leggermente maggiori per quanto riguarda FNR, precisione e sensibilità (ad es: 0.17,

0.12 e 0.17 per quanto riguarda Gradient boosting). Per via di questa variabilità intrinseca, la performance di classificazione del singolo modello ottimizzato può essere sostanzialmente differente dalla performance media. Ulteriori studi su un campione più esteso potrebbero permettere di minimizzare questa variabilità, riducendo la dipendenza dai dati, in quanto ogni partizione tende a essere ugualmente rappresentativa del dataset originale quando la dimensione del campione viene notevolmente aumentata. In seguito, si potrebbero definire dei modelli ottimizzati, e si potrebbe ottenere una regolazione più fine degli iper-parametri in modo da avvicinarsi ulteriormente allo sviluppo di un'applicazione smartphone di screening uditivo per adulti.

Sono necessarie ulteriori ricerche per validare il test di speech-in-noise e il metodo di classificazione sia per il criterio 1 che per il criterio 2, su una popolazione molto più ampia che coinvolga un numero maggiore di soggetti ipoacusici, anche con gradi di ipoacusia più severi.

Keywords: perdita uditiva; screening uditivo; audiometria tonale; test speech-in-noise; apprendimento supervisionato; classificazione; machine learning; modelli lineari generalizzati

Contents

List of Figures

List of Tables

1. Introduction

1.1 The auditory system

The auditory system consists of three main components: the outer, the middle, and the inner ear, allowing sounds to be transferred from the external environment to the high processing centers in the brain.

A sound is defined as a mechanical vibration (i.e., alternation of rarefaction and compression of the molecules) propagating through an elastic medium.

The general structure of the auditory system is represented in Figure 1.

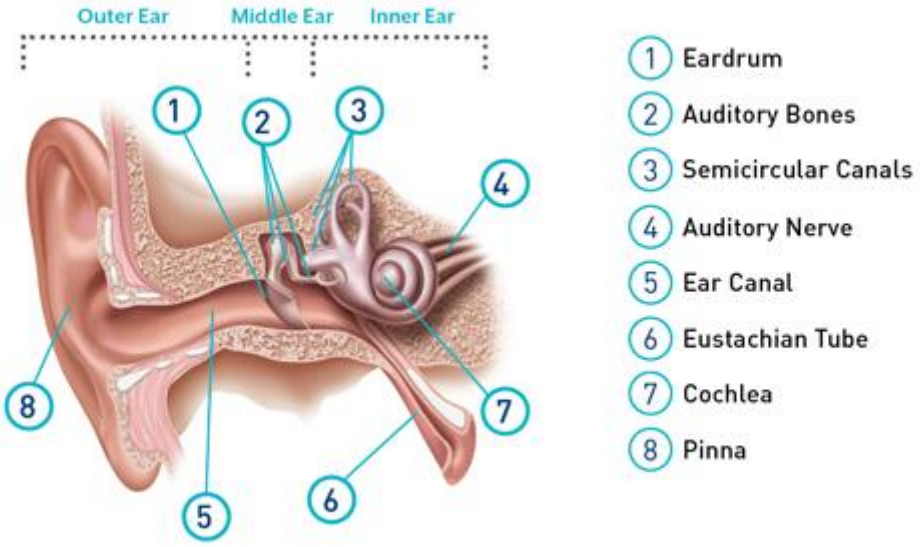


Figure 1. Main components of the auditory system and the sound transmission chain.

The outer ear comprises the pinna, also called auricle, which is the external cartilaginous structure that collects sounds and funnels them inward and the ear canal, located in the temporal bone, that acts as a resonator with a peak at 3kHz and allows sounds to travel toward the eardrum.

The middle ear includes the eardrum, or tympanic membrane, which is a thin layer of tissue that vibrates when a pressure wave hits it and the ossicular chain, formed by 3 small bones called malleus, incus and stapes, that amplifies the signal received from the eardrum and transfers energy to the inner ear that, in turn, allows the motion of the cochlear fluid. Since the impedance of the air is much lower with respect to the one of the internal fluids, the middle ear act as a transformer, ensuring impedance matching.

Finally, the inner ear is composed by the cochlea, which is the hearing sensory organ and the vestibular system that controls equilibrium and balance of the human body.

The cochlea is a snail-shaped structure that comprises three chambers filled with fluids (Scala vestibuli, Scala tympani and Scala media) and contains the basilar membrane and the Organ of Corti. The basilar membrane works as a bank oscillator and exploits the concept of frequency coding, separating sounds according to their frequency. Indeed, each section of the basilar membrane presents a maximum oscillation at a certain frequency. The basal portion of the basilar membrane encodes high frequencies whereas the apex encodes lower frequencies.

The Organ of Corti is characterized by the presence of ciliated cells, called hair cells, arranged in rows along the basilar membrane. Hair cells are the sensory cells that convert the vibration of the membrane into a 'code' understandable at neural level. These cells have bundles of stereocilia on their apical surface and are divided into inner and outer hair cells, exploiting different functions, along with different structures. Inner hair cells act as transducers of a sound in the audible range in an afferent nervous impulse. In particular, the oscillation of the basilar membrane causes a deflection of the stereocilia which put the cross link under tension and in turn causes the opening of ionic channels in the cellular membrane. The incoming external sound is therefore converted into pulses delivered to the auditory nerve (VIII cranial nerve). The pulses are then carried to the brainstem and finally processed in the nuclei of the temporal lobes.

Alteration and damages, as well as pathologies, can afflict the different blocks of the auditory chain causing a loss of the subject's hearing ability.

1.2 Hearing loss

As stated by the World Health Organization, it is estimated that over 6.1% of the world's population suffers from hearing loss, making it the fourth leading cause of disability worldwide (World Health Organization, 2018). Almost half a billion people are affected by this disability, of which 93% are adults and 7% children, and this number is expected to double by 2050.

A normal hearing person is defined as having hearing thresholds of 25 dB or better (i.e., lower) in both ears. People with worse hearing abilities than these have a disabling hearing

loss. Hearing loss can have different degree of severity (i.e., mild, moderate, severe, or profound) and may affect only one ear or both.

Mild to severe hearing loss is referred to as 'Hard of hearing' and leads to the need of assistive devices such as hearing aids or cochlear implants. People belonging to these hearing loss categories can communicate with spoken language and do not need to use sign language, which is instead a necessity for people with profound hearing loss, referred to as deaf people.

Hearing loss can be due to different causes, which are grouped into two main families: congenital or acquired. Congenital causes are related to hearing loss already present at birth or manifested slightly after and may be due to genetic factors, both hereditary and non-hereditary, complications during pregnancy (infections, inappropriate use of particular drugs..) or immediately after birth (i.e., low weight, asphyxia…). Acquired causes, instead, can occur at any age and some examples are meningitis and other infections, otitis media (fluid accumulates in the ear, quite commonly among children), head injuries, excessive noise (i.e., occupational noise or prolonged exposure to loud sounds for entertaining purposes).

According to the section of the auditory system affected, 3 kinds of hearing loss can be distinguished: conductive hearing loss, sensorineural hearing loss and mixed hearing loss.

Conductive hearing loss is due to damages or obstructions within the outer or middle ear, that prevent sounds from reaching the inner ear properly. This kind of hearing loss, which has no frequency dependency, results in a flat audiogram.

Sensorineural hearing loss, instead, is more common than the previous one and it is due to damages to the inner ear nerves and cells, which prevent sound from being projected at higher levels. In this type of hearing loss, some frequencies are usually more involved than others (typically LF in babies and HF in older people). This disease hugely affects hearing and, in most cases, cannot be resolved by surgical procedures, however the use of hearing aids can be helpful.

Lastly, when the damage involves both the outer and inner ear or the auditory nerve, so there is a combination of conductive and sensorineural hearing impairment, this refers to a mixed hearing loss. Figure 2 shows the main causes leading to different types of hearing loss.

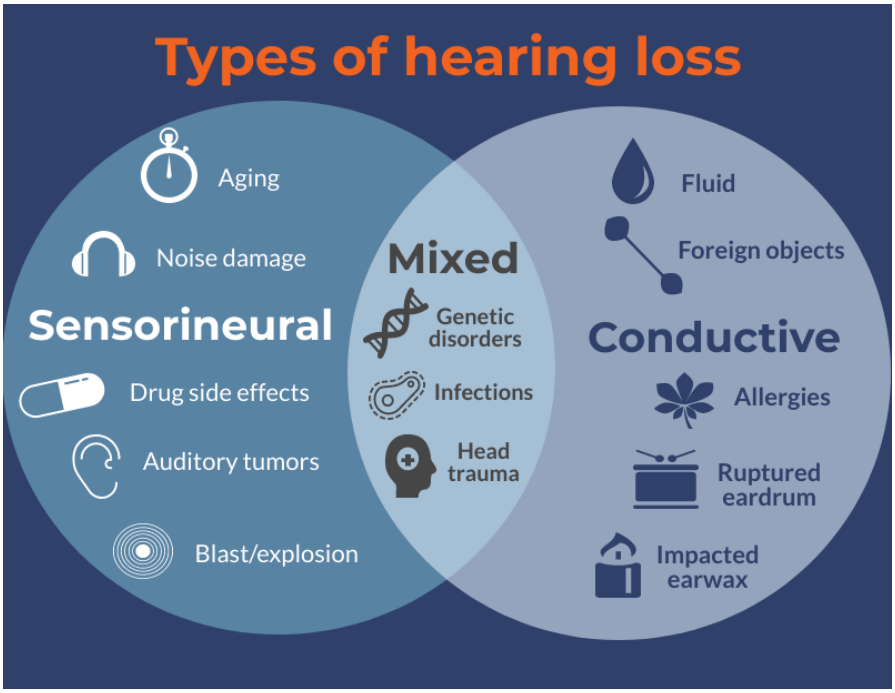


Figure 2. Main leading causes of sensorineural, conductive, and mixed hearing loss.

Hearing impairment can have a very strong social, functional and emotional impact on the affected person, mainly because it affects the ability to communicate with other people; in fact it can lead to educational problem such as delays in the development of communication skills, poor academic performance and need for educational assistance, social isolation and difficulty in finding a job. Most of these problems can be mitigated by early detection and the use of assistive technologies such as cochlear implants.

According to WHO, almost 50% of hearing loss can be prevented by specific strategies (limit exposure to loud sounds, wearing earplugs…) or by screening campaign and early interventions. However, routine medical examinations for adults usually do not include hearing tests and individuals manifesting symptoms of hearing loss tend to seek help very late and underestimate the problem, as a loss of hearing capacities with age is commonly considered inevitable (Paglialonga, et al., 2020). A widespread diffusion of hearing screening can help overcome this problem. Indeed, encouraging periodic screening initiatives at school and in the workplace, but also the distribution of smartphone apps for hearing screening, can be a useful tool to promote awareness and early identification of hearing loss.

1.3 Hearing tests

Different types of hearing tests can be identified depending on the age of the patient i.e. newborns, children, and adults. The main standard hearing tests are listed below:

– *Pure-tone threshold audiometry (PTA)* is the gold standard for audiologic examinations, based on the measurement of hearing sensitivity to pure tones, including both air-conduction measurements (via headphones) and bone-conduction measurements (via oscillators placed on the mastoid bone or the forehead).

  Hearing test results are synthetized in a graph called audiogram, with the x-axis representing the tested frequency and the y-axis representing the hearing threshold. The plotted graph includes all hearing thresholds, measured in decibel, for all tested frequencies, for both ears, as it is shown in Figure 3.

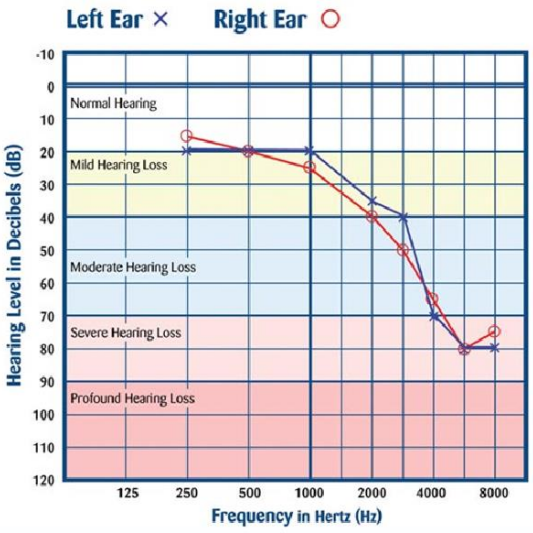


Figure 3. Example of an audiogram, showing the hearing thresholds at the tested frequencies (250-8000 Hz), for both ears.

Normally, the curve related to bone conduction has quite lower (better) values with respect to the air conduction one. The difference of air conduction curve and bone conduction curve in the audiogram is referred to as air-bone gap.

In normal-hearing subjects, hearing thresholds are lower than 25 dB. In presence of sensorineural hearing loss there is no air-bone gap and the curves related to air and bone conduction present higher (worst) values with respect to the normal range.

In conductive hearing loss instead, the bone conduction curve remains unchanged while the hearing thresholds related to air conduction worsen, usually showing some frequency dependency. Finally, in mixed hearing loss, the audiogram presents a worsening in both air and bone conduction trends but also an air-bone gap is present.

- *Tympanometry* is a test to assess functioning of the middle ear, especially the Eustachian tube, and the eardrum. The test consists in evaluating the compliance of the eardrum by sending air pressure, perceived as a tone, into the ear and recording the consequent movement of the eardrum. The graph summarizing the results, called tympanogram, shows the movement of the eardrum in response to the air pressure created by the tympanometer. The classical shape of the graph can change; for instance, in presence of perforation of the eardrum, since there is no more movement of the membrane, it presents a flat line.
- *Auditory Brainstem Response (ABR)* is an example of far-field sound-evoked electrical potential that tests the functioning of the inner ear and brain auditory pathways by placing electrodes on the scalp and recording brain activity in response to the sound delivered. The measurement of this evoked potential produces seven vertex positive waves marked with roman numbers, each representing the activity of a particular branch of the auditory path from the cochlea to the upper levels. ABRs can be typically observed in a window of 10-20 ms after the presentation of an acoustic stimulus (delivered by a specific stimulating system with a fixed repetition rate) and relevant parameters are latencies and amplitudes.
- *Otoacoustic emission (OAE*) presence is widely considered in hearing screening programs to assess cochlear functioning especially in newborns and children, to help the diagnosis of frequency specific hearing loss or monitoring noise exposure effect or consequences of exposure to ototoxic agents. An OAE is a low-level sound emitted by the cochlea that can be both produced spontaneously or evoked by a specific acoustic stimulus and is measured in the outer auditory canal. The presence of OAEs can provide information about the correct functioning of the auditory conductive mechanism and can be related to the mobility of the outer hair cells, however the OAE itself is very subject-dependent.

  OAEs are absent or very weak in presence of hearing losses.

1.4 Speech-in-noise tests

Pure-tone audiometry is considered the gold standard for audiological screening, even though it is not exempt from some limitations including for instance the calibration of the transducers, the need for a low noise (or sound treated) environment and the need of well-trained health care operators available. Moreover, PTA requires a certain amount of time (more than 20 minutes) to set up the instrument, calibrate it and perform the procedure (Jansen, Luts, Dejonckere, van Wieringen, & Wouters, 2013). The combination of these factors contributes to the fact that PTA is not suitable for large scale or remote screening procedures that perhaps could be performed in an automated, quicker way, over the Internet, using a domestic audio equipment but still reaching the same goal which is to promote awareness of the individual hearing conditions and early detection of hearing loss.

Furthermore, routine audiological tests may not be sufficient to highlight problems related to speech in noise comprehension. Indeed, poorer speech recognition abilities in noise in the presence of an audiogram presenting normal threshold values have often been reported. These are referred to as 'hidden hearing loss' as they cannot be detected by standard hearing tests. Reduction of speech understanding, especially in noise, was found to be one of the prevailing symptoms of noise-induced hearing loss (NIHL) and age-related hearing loss (presbycusis), which are the most common forms of hearing loss (Van Eynde, Denys, Desloovere, Wouters, & Verhaert, 2016).

Speech-in-noise tests (SNTs) can be considered valuable extensions of standard clinical hearing tests, improving diagnosis in presence of the so called 'hidden hearing loss' but also allowing a quicker, automated and remote way for hearing screening. Usually, the main outcome of a speech-in-noise test is the speech reception threshold (SRT), defined as the signal-to-noise ratio (SNR) at a certain intelligibility level.

Several studies have proved a reasonably high correlation between SRT and the hearing thresholds at various frequencies as provided by pure tone audiometry (Smoorenburg, 1992) (Jansen, Luts, Dejonckere, van Wieringen, & Wouters, 2013). Thus, it can be reasonably deduced that SRT could be a key feature for PTA outcomes prediction.

Several speech-in-noise tests have been implemented and validated, differing on provided stimuli (i.e., speech and noise), test procedures and target age; some examples are briefly described hereafter.

The Quick Speech-in-noise test (QuickSIN) evaluates the ability to recognize speech in noise using 6 Institute of Electrical and Electronics Engineers (IEEE) sentences, extracted from 12 lists, in the presence of four-talker babble, maintaining a fixed level for the speech material and a noise level increasing in 5 dB steps between stimuli proposal, indeed the signal-to-noise ratio decrease by the same amount, increasing the difficulty of the task at each trial. The repetition of the 5 keywords included in each sentence provides an incrementation of the test score and the final result is reported in terms of SNR required to correctly identify 50% of the proposed stimuli (Holder, Levin, & Gifford, 2018).

The Dutch National Hearing Test (NHT) is the precursor of adaptive digit-in-noise tests, in which several triplets of digits with a background masking noise are delivered via telephone and the participant has to type on the keypad the sequence heard; if the sequence is correct, the level of the speech material decreases of 2 dB, vice-versa, after a wrong sequence typing, the level increases of the same amount. The noise level, instead, remains fixed for the whole test execution. The SRT is finally calculated as the average SNR of the last 20 trials, with the first four trials being discarded to allow familiarization with the test (Smits, Kapteyn, & and Houtgast, 2004). This test was subsequently adapted and validated for several languages and also computer versions were developed (Folmer, et al., 2017).

Another application of online speech-in-noise test for audiologic screening is Earcheck (EC), a test based on a set of nine Consonant-Vowel-Consonant (CVC) syllables, selected from the Dutch wordlist used for speech audiometry with diagnostic purpose. For each proposal of the stimulus, all nine possible responses are displayed, together with an option ‘not recognized’ to be selected whenever the user does not identify the stimulus. SRT is then calculated by averaging the SNR values of the trials from 7 to 27. Finally, by the definition of 3 cut-off values (-10 dB, -7 dB and -4 dB), appropriately chosen (Martens, Perenboom, Ploeg, & van der Dreschler, 2005), the results are categorized into 4 possible status, respectively good, moderate, insufficient and poor hearing level.

Subsequently, a variant of Earcheck, called Occupational Earcheck (OEC), specifically addressed to detect high-frequency hearing loss due to occupational reasons has been developed; in this application the CVC words are chosen in a way to contain a higher proportion of high frequency consonants.

The Speech Understanding in Noise test (SUN) (Paglialonga, Tognola, & Grandori, 2014) is another striking example of automated, speech-in-noise test based on the identification of

meaningless Vowel–Consonant–Vowel (VCV) sequences in presence of a background noise where stimuli have to be recognized among three possible alternatives (3AFC task). The output level of the speech material is fixed to a level of 60 dB HL whereas the signal-to-noise ratio, and consequently the noise, is progressively varied from a SNR of +8 dB to a ending value usually comprised between -6 and -4 dB. The final score is computed as the number of correctly discriminated stimuli with respect to the total number of proposed stimuli and associated with a short statement that warns the subject whether a hearing check is suggested.

SNT tests have been largely investigated in the last few decades and many examples are available online, providing an accessible, fast, easy to use, costless and automated way to perform reliable screening tests. However, the diffusion on a large scale of the majority of these tests is prevented by the fact that they typically use sentences, words or digits as speech material, therefore they are fully language dependent. As a consequence, each new version, before being diffused in a country with different native language, has to be translated, adapted and validated (Rocco, 2018). On the other hand, a test like this, proposed in only one language, could lead to a smaller pool of users, subsequent inequalities in the access to screening depending on the native language and possible inaccurate results. Consequently, trying to reduce the language dependency is one of the main design requirements to consider while developing a SNT available as a smartphone app with the purpose to achieve widespread screening. In this context, the replacement of sentences and words with meaningless Vowel-Consonant-Vowel syllables has been proved to be helpful in achieving stable performance on listeners of unknown and various languages (Paglialonga, et al., 2020).

1.5 Machine learning for medical applications

Machine learning (ML) is a broad field of artificial intelligence that involves the design of computational algorithms learned by experience i.e. starting from large amounts of available data, in order to make decisions and predict new data.

Learning approaches are data-driven, since the efficiency and accuracy of their results largely depend on the data initially considered. Hence, this discipline must be associated also to data analysis, probability and statistical concepts (Mohri, Rostamizadeh, & Talwalkar, 2018). A learning task can be supervised or unsupervised; supervised learning focuses in

learning a function that relates in the best possible way input samples to their correspondent known outputs (targets) and it is exploited in classification and regression problems, whereas unsupervised learning (e.g. clustering) tries to learn the natural structure of the data points (recurrent patters or affinities), without using any explicit label.

Classification, which is the most common supervised learning task, implies mapping input data with discrete output labels (classes) whilst regression relates input data to a continuous output.

ML includes a wide range of practical applications, from text classification, to computer vision (e.g. face detection) and computational biology and, last but not least, medical applications. Machine learning approaches have been proved to be very helpful in assisting medical diagnosis (e.g. diabetic retinopathy) and identifying first stages of appearance of the pathological condition. Screening phase and early diagnosis are extremely important to set the basis for the medical decision-making process and the choice of a future treatment to follow. Besides, healthcare does not have enough medical experts to cover the huge amount of screening tests that should be done in order to protect the global population from severe diseases. Perhaps, screening applications involving artificial intelligence can be used to automate a variety of screening process and moreover to help medical decisions (Kumar, 2018). In fact, ML algorithms may be able to detect particular patterns of data (e.g. biological factors, social factors, measurements) that behave as early indicators for pathologies and that would not be noticed by a human eye.

Deep learning has also opened the door to innovation in audiology, for instance, researchers have focused on hearing aid personalization (also referred to as hearing aid fitting).

In most hearing aids, gain and compression settings are defined from the patient's audiogram and then adjusted, however, each different user is exposed to unique auditory scenarios and varying auditory intentions and therefore requires a number of specific custom settings to improve the listening experience. The recently developed WIDEX EVOKE™ device exploits machine learning techniques to allow automatic adjustment and optimization of settings (Townend, Nielsen, & Jesper, 2018) of hearing aids in different listening situations, based on what has been learned from the user's past experience.

Another example of application in the audiological field is a recent retrospective study and comparison of different machine learning algorithms (Random Forest, Support Vector Machine, Multilayer Perceptron, K-Nearest Neighbor and AdaBoost) for the prediction of

the prognosis in sudden sensorineural hearing loss, starting from a set of 31 and 15 prognostic factors extracted from demographic data, medical records, inner-ear symptoms, pure-tone audiometry, and laboratory data (Park, et al., 2020).

Finally, a last example that is worth mentioning concerns the use of ML techniques (Support Vector Machine, Neural Network, Multilayer Perceptron, Random Forest, and Adaptive Boosting) for the prediction of NIHL (noise induced hearing loss), defined as the average value of the hearing thresholds at 1000, 2000, 3000 and 4000 Hz exceeding 25 dB HL, in noise exposed workers, based on a set of attributes including age and duration of exposure. (Zhao, et al., 2019)

1.6 Thesis objectives

The aim of this study was to address specific machine learning algorithms to investigate the ability of a novel automated, self-operated and fast speech-in-noise test recently developed by Politecnico di Milano in collaboration with Consiglio Nazionale delle Ricerche (CNR), to identify individuals with slight/mild or higher degree of hearing loss,.

The goal was to investigate whether the test might be suitable and accurate enough to be validated as an adult hearing screening tool for remote use through the future implementation of a smartphone app.

SRT, which is the usual primary outcome of speech-in-noise tests, has already been proven to be a good predictor for PTA results on several occasions (Zanet, Polo, Rocco, Paglialonga, & Barbieri, 2019) (Paglialonga, et al., 2020). The study aims to analyze, not only the SRT, but also other features extracted during the experimental procedure, to investigate what may be the optimal combination of variables to detect, with a good accuracy, the presence of hearing loss, according to the WHO criteria for hearing impairment, specifically based on the average pure-tone hearing threshold computed on the four central frequencies (0.5, 1, 2, and 4 kHz).

Based on the results collected during a series of hearing screening initiatives at local level, a dataset with 156 records has been created. The screening result, to be used as label for supervised learning techniques, was set to 0 (‘pass’) or 1 (‘fail’) according to two of the WHO criteria for hearing impairment: PTA> 25 dB HL for mild hearing impairment, (criterion 1), and PTA> 40 dB HL for moderate hearing impairment (criterion 2).

First, the possible influence of sex, ear tested (left/right), and age on the collected sample was investigated.

As a second step, the distribution of the variables and their correlation have been analyzed.

Then, generalized linear models (GLMs) were addressed to investigate the ability of each single feature to predict PTA outcomes and spot a subset of most suitable features for classification.

Finally, seven among the most used classification methods (Decision Trees, Support Vector Machines, logistic regression, K-Nearest-Neighbors, ensemble logistic regression, Random Forest and Gradient Boosting) have been implemented, using seven different sub-sets of variables appropriately chosen and their performance in detecting hearing loss has been compared by means of different metrics.

Once the most suitable methods have been highlighted, the classification performance of the new test has been compared with that of some of the most popular SNTs.

2. Background

2.1 Test design and implementation

Recently, a new automated English-based SNT test for adult hearing screening has been developed (Polo & Zanet, 2018). Both test and GUI were implemented using MATLAB (R2017a, Mathworks™). Accuracy in estimating SRT with respect to conventional staircase procedures, reliability and test-retest repeatability were subsequently investigated on a sample composed of 26 normal hearing non-native young adults and a second group of 72 unscreened adults and older adults with various native languages, showing varying degrees of hearing impairment (Paglialonga, et al., 2020).

The main features of the test are here listed and discussed (Polo, et al., 2020):

- *Automated, self-operated, with a user-friendly graphical interface, optimized for the use with touchscreens*.
- *Usable for subjects with unknown language and different levels of education*. The selected speech material is based on Vowel-Consonant-Vowel (VCV). As VCVs (which will be further discussed in the next section) are independent on semantics, no effort has to be done in order to understand the meaning of the proposed stimuli. Moreover, the use of meaningless VCVs, together with the use of a three-alternative forced-choice (3AFC) task where the proposal of the three possible alternatives is based on a maximal opposition criterion (i.e., the consonants of the two wrong VCVs differ from the listened one in terms of manner, voicing and place of articulation), guarantees a limited involvement of higher level processing centers, limiting the influence of education and native language on test results and allowing the fruition of the test to a larger population, without inequalities and penalizations in accessing the screening procedure (Cooke, Lecumberri, Scharenborg, & Van Dommelen, 2010). Preliminary studies have highlighted a stable performance in subjects of different languages. However, further investigation on a larger population spanning different languages is needed. (Paglialonga, et al., 2020)
- *Quick assessments of speech recognition in noise.* The test lasts on average 3 minutes and 30 seconds, about 2 minutes shorter than the conventional staircase procedure, for both subject with normal hearing (Zanet, Polo, Rocco, Paglialonga, & Barbieri, 2019) and subjects with hearing loss (Paglialonga, et al., 2020). The reduction in time

is due to having abandoned the assumption that the intelligibility of the proposed stimuli is homogeneous, together with the use of an optimized novel staircase procedure with upward and downward steps adaptively based on the estimated psychometric curve of the stimuli, selecting an optimal ratio between the steps of 0.74. These solutions have allowed a faster convergence and the need of only 12 reversals to complete the test execution, ensuring that the test is completed in less time with respect to conventional procedures based on the use of fixed, equal upward and downward steps, converging after a certain number of reversals (i.e., after 20 reversals for a standard one-up/three down staircase), when the probability of a decrease in the presentation level matches the probability of an increase in the presentation level.

- *Reduced intra-individual variability.* Test-retest experiments have been carried out to verify the intra-individual repeatability of test results. The newly developed SNT test provides repeatable results in terms of SRT and performance (number of presented stimuli, test execution time, and percentage of correct responses) for 98 individuals with different degrees of hearing impairment. Indeed, the average change in SRT between test and retest trial was not statistically significant (-0.35 dB) (Paglialonga, et al., 2020) (Zanet, Polo, Rocco, Paglialonga, & Barbieri, 2019). Furthermore, thanks to the test design choices, no significant learning effect was observed between the first and the second trial, whereas usually an improvement in the result is noticeable in subsequent trials because the subject is more used to the interface and the way stimuli are proposed, becoming more able to distinguish speech components from noise (Leensen, de Laat, & Dreschler, 2011).
- *Able to provide accurate screening results, identifying hearing impairment, if any.* The test classification performance based on SRT only (i.e., ability to discriminate subjects with PTA higher than 25 dB HL at the four central frequencies), on a population of 98 subjects, yielded a moderate accuracy (82%) and an area under the ROC curve (AUC) equal to 0.84 (Paglialonga, et al., 2020), comparable with other SNTs based on similar approaches in terms of type of stimuli, cut-off criteria and testing procedure. Further analysis, on a larger sample (148 subjects) and including multivariate classification algorithms, will be described in the next chapters.

– *Viable for remote testing in uncontrolled environmental conditions and with unknown transducers.* The proposed test assured reliable SRT estimates and test-retest repeatability both considering output levels controlled by the audiometer (controlled environmental noise settings) and allowing self-adjustments of the test volume before the test starts (uncontrolled settings), on a group of 26 normal hearing young adults (Zanet, Polo, Rocco, Paglialonga, & Barbieri, 2019). Moreover, a quantitative analysis of the influence of different transducers on the test output levels have been recently carried out on a larger population, revealing that earphones yielded generally lower output SPL with respect to headphones; therefore, the choice to include the option to self-adjust the volume to a comfortable level based on the subject perception may partly compensate differences due to the chosen transducer. However, headphone use is recommended because the maximum SPL reached by earphones (at least for the models tested) might not provide enough intelligible speech stimuli to subjects showing reduced hearing sensitivity and further analysis on a larger number of transducers is needed (Polo, et al., 2020).

2.2 Test stimuli

English has been chosen as the language for the test as it is the most suitable language for the development of a language-independent speech-in-noise test, showing the best outcomes in terms of intelligibility, considering 6 languages among the 20 languages with at least 50 million mother-tongue speakers (i.e., English, French, Italian, German, Portuguese and Spanish) (Rocco, 2018).

The provided set of stimuli includes 12 Vowel-Consonant-Vowels (VCVs) composed by consonants in the context of the vowel ‘a’ (i.e., aba, ada, afa, aga, aka, ala, ama, ana, apa, ara, asa, ata). Stimuli presented as intervocalic consonants may be helpful in adult hearing screening applications because one of the first hint of hearing loss related to aging is the manifestation of a decreased ability in consonant recognition. (Killion & Niquette, 2000).

All the VCVs were recorded in a sound-treated room by a professional male native English speaker, pronounced without prosodic accent and with constant pitch.

The recording phase took place in a professional recording studio by means of a Neumann TLM 103 microphone, a SSL S4000 64- channels mixer, Motu HD 192 A/D converters (44,1

kHz, 16 bit), and GENELEC 1025A control room monitor (Vaez, Desgualdo-Pereira, & Paglialonga, 2014) (Paglialonga, Tognola, & Grandori, 2014).

The recorded VCVs were digitally equalized to fulfill the equal speech level requirement as defined by ISO 8253-3:2012 Standard. The noise added to the signal has been generated by filtering a Gaussian noise of amplitude equal to the average level of the speech material (i.e., the VCVs) by the International long term average speech spectrum (LTASS) (Byrne, Dillon, & Tran, 1994) and a low pass filter (cut-off 1.4 kHz, roll-off slope 100 dB/octave) and finally by adding a noise floor (i.e., the same filtered noise attenuated by 15 dB) (Leensen, de Laat, & Dreschler, 2011).

Among the 12 selected VCVs, significant differences in terms of intelligibility has been found. Therefore, 4 clusters of VCVs with homogeneous intelligibility have been created.

2.3 Experimental procedure

The experimental procedure consisted of three phases: a preliminary Pure Tone Audiometry assessment, the compilation of personal data and of the HHIE-S questionnaire and finally the submission to the Speech-in-noise test.

The experimental protocol was approved by the Research Ethical Committee of Politecnico di Milano (Opinion n.2/2019). All participants took part to the study on a voluntary basis and before being tested, signed an informed consent. All the data were used for research purposes only.

Testing procedure was briefly explained to every subject before the session started. Participants were asked to turn off their mobile phones and keep answering even if they were not aware of the right answer. The test was stopped and started every time there was an intense and sudden environmental noise, to prevent it from affecting the experimental results.

PTA was assessed by means of a clinical audiometer (Amplaid 177+, Amplifon TM), calibrated using a 1 kHz sinusoidal wave as suggested by ISO 8253-1:2010.

Pulsed 2 Hz waves were used as stimuli, tested frequencies were in the range 250 Hz-8000 Hz. The potential influence on the outcomes due to uncontrolled conditions was tested. In particular, the volume was set by the participants themselves through a slider in the GUI.

PTA procedure was carried out according to the audiometer manual guide (Amplaid a137-a177 plus instructions manual), following the steps here described:

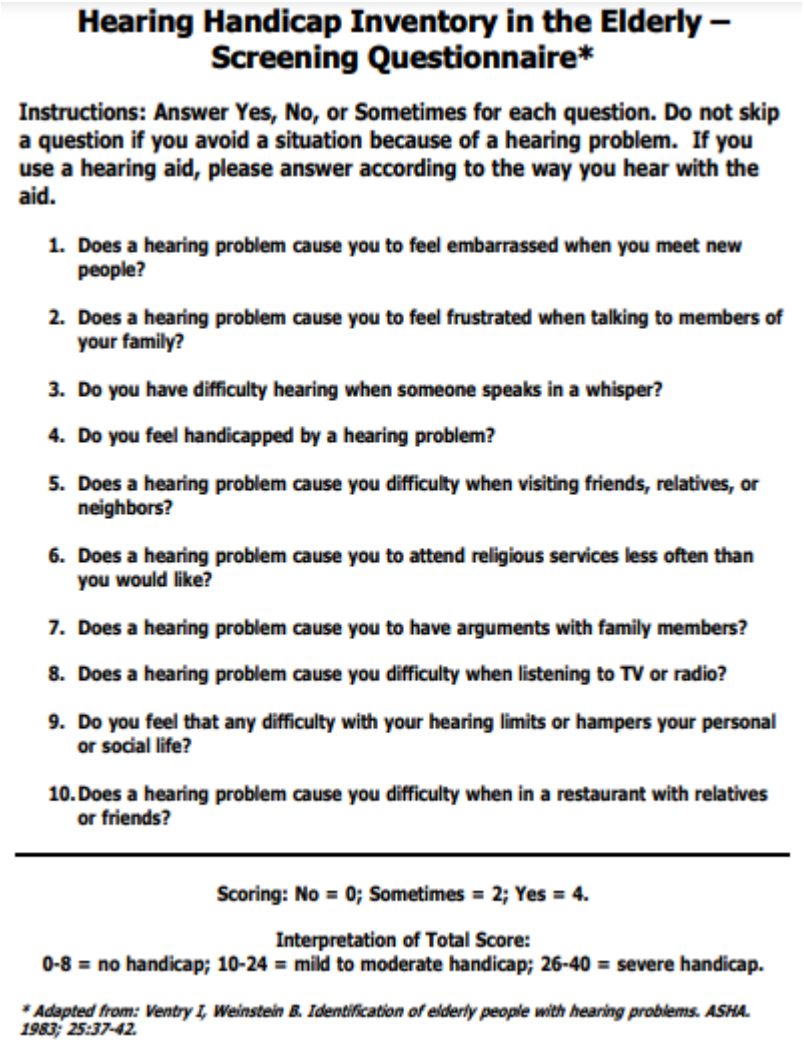
I. The subject was asked to raise a hand as soon as he heard a tone and keep it raised until the stimulus was perceived. The participant should not be able to see the operator changing the parameters of the audiometer.
II. Pure tones (i.e., single-frequency tones) were produced by the oscillator and delivered at different frequency values, following this sequence: 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, 8000 Hz, 1000 Hz, 500 Hz and 250 Hz.
III. To ensure the familiarity of the participant with the task, the initial tone had to be clearly audible, with an intensity level that was high enough to be easily perceived.
IV. Each time the tone was perceived, the presentation intensity level was reduced by 10 dB. On the contrary, if no response occurred, the stimulus level was increased by 5 dB until the subject raised its hand (i.e., perceived the stimulus) and then decreased by 5 dB until the participant couldn't perceive the stimulus anymore. The last perceived stimulus was the correspondent hearing threshold at the frequency considered.
V. Point 3 and 4 were then repeated for all the other frequencies.

The pure-tone threshold is defined as the lowest audible intensity level of a pure tone that evokes a response 50% of the time. However, in practice, it is estimated as the intensity level that induces a response in 100% of the cases.

After PTA assessment, which lasted at least 20 minutes, the subject compiled the ten questions of the Hearing Handicap Inventory for the Elderly Screening Version (HHIE-S). HHIE-S is a questionnaire that allows to have a clue about self-perceived hearing conditions of the subject by asking questions related to real-life situations. Indeed, it investigates perceived presence of functional limitation (e.g. social, emotional issues) related to possible hearing impairment conditions (Tomioka, et al., 2013).

The structure of the English version of the questionnaire is reported in Figure 4; during the experimental procedure, the questionnaire was presented to each subject in their native language. Each individual question has three possible answers, with an associated score: yes (4 points), sometimes (2 points) or no (0 points). The final score of the questionnaire is the

sum of the points obtained for each answer. Hence, the maximum achievable score is 40 (maximum hearing handicap).



Hearing Handicap Inventory in the Elderly – Screening Questionnaire*

Instructions: Answer Yes, No, or Sometimes for each question. Do not skip a question if you avoid a situation because of a hearing problem. If you use a hearing aid, please answer according to the way you hear with the aid.

1. **Does a hearing problem cause you to feel embarrassed when you meet new people?**
2. **Does a hearing problem cause you to feel frustrated when talking to members of your family?**
3. **Do you have difficulty hearing when someone speaks in a whisper?**
4. **Do you feel handicapped by a hearing problem?**
5. **Does a hearing problem cause you difficulty when visiting friends, relatives, or neighbors?**
6. **Does a hearing problem cause you to attend religious services less often than you would like?**
7. **Does a hearing problem cause you to have arguments with family members?**
8. **Does a hearing problem cause you difficulty when listening to TV or radio?**
9. **Do you feel that any difficulty with your hearing limits or hampers your personal or social life?**
10. **Does a hearing problem cause you difficulty when in a restaurant with relatives or friends?**

Scoring: No = 0; Sometimes = 2; Yes = 4.

Interpretation of Total Score:
0-8 = no handicap; 10-24 = mild to moderate handicap; 26-40 = severe handicap.

** Adapted from: Ventry I, Weinstein B. Identification of elderly people with hearing problems. ASHA. 1983; 25:37-42.*

Figure 4. Hearing Handicap Inventory for the Elderly Screening Version (HHIE-S) questionnaire, including 10 questions aimed at the self-perception of social and emotional consequence of hearing loss.

Unlike the first study on a population of 26 YA (Polo & Zanet, 2018), the audiometer has no longer been used to control the output volume of the VCV stimuli. In fact, the subject could adjust the volume through a slider of the developed GUI right before the test execution. After selecting the volume and the ear to be screened, the test could start.

At each trial, one acoustic stimulus in the form of VCV is delivered and the user had to choose among three possible alternatives (3AFC, three-alternative forced choice). The test

exploits the concept of cross-cluster staircase, where intelligibility steps cross psychometric curves (i.e., identifying the four homogeneous clusters of VCVs), instead of the traditional staircase, based on the assumption of having an homogeneous set of stimuli and therefore a single psychometric curve. The first stimulus is delivered at high values, with an intelligibility of 98.34 % and a SNR of 8 dB, so to initially facilitate the comprehension and allow the subject to gain familiarity with the test. Then, a 1-up/3-down (1U3D) method is adopted, indeed intelligibility of the proposed stimulus is increased after each wrong answer and decreased after three right detections of the VCVs.

Each step along the psychometric curve is defined by two coordinates: intelligibility and SNR, with step size varying along the curve (i.e., bigger steps in the middle range, characterized by a quite linear trend and smaller steps for low and high values of SRT and intelligibility, where non-linearities are prevalent). The ratio between upward and downward steps is fixed and equal to 0.74.

The test stops after 12 reversal (i.e., inversion of the direction). A reversal is obtained every time an increase in the stimulus presentation level is followed by a decrease of the same or vice versa. The first reversals (familiarization phase) are discarded so to avoid possible mistakes due to the subject getting used to the test. After the end of the test, a summary of the results is displayed, including test duration, number and percentage of correct responses, the intelligibility threshold and the speech reception threshold (SRT), together with a graphical representation of the cross-cluster staircase.

The SRT (i.e. the signal-to-noise ratio at a certain level of intelligibility) is defined by two coordinates: intelligibility and SNR. The first coordinate is estimated as the mean of the last ascending runs of the cross-intelligibility staircase, whereas the second coordinate is found by the intersection with the average psychometric curve.

An example of the final test screen is shown in Figure 5.

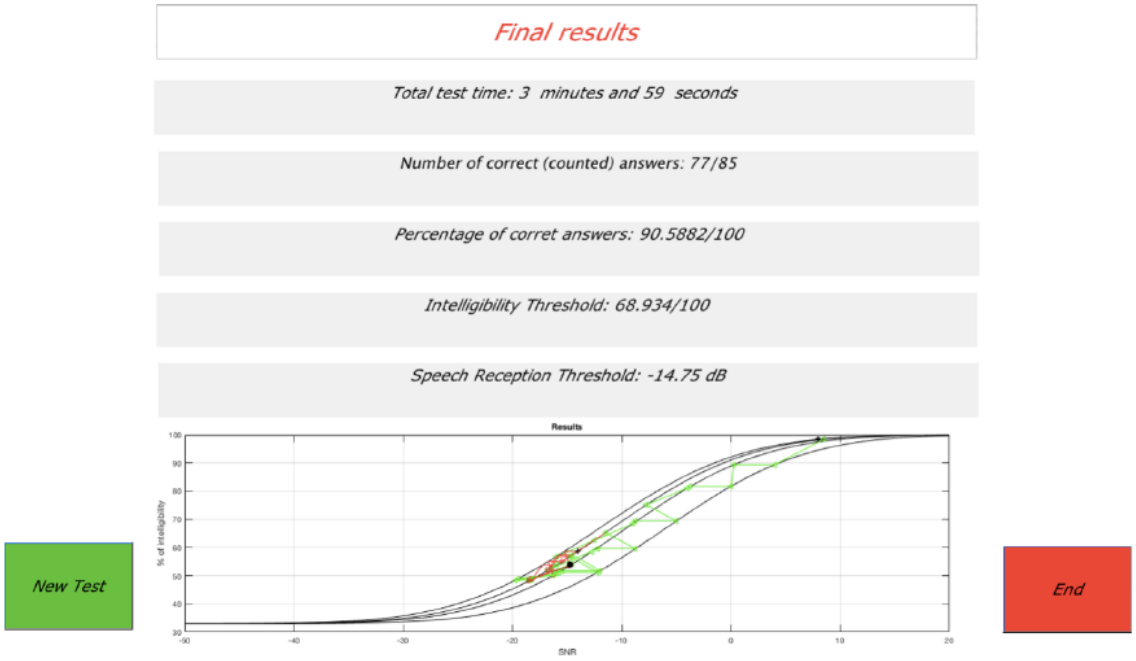


Figure 5. Sample screenshot of test outcomes. A summary of the main information extracted is displayed. At the bottom, a graphical representation of the cross-cluster staircase is shown. Correct trials correspond to green arrows whereas incorrect trials are represented as red arrows. The black point stands for the SRT value, computed along an Average Psychometric Function.

3. Materials and methods

3.1 Participants

Participants finally considered were 148 adults (age = 52.1 ± 20.4 years; age range: 20-89 years; 46 male, 102 female) of different native languages (Italian: 118 subjects; English: 10 subjects; Arabic: 6 subjects; Spanish: 4 subjects; French, Somalian: 2 subjects; Albanese, Filipino, German, Moroccan, Igbo and Efik: 1 subject) tested in uncontrolled environmental noise settings in the laboratory and at local health screening initiatives (i.e., at senior citizens' universities, health prevention and awareness events for the general public). Given the opportunistic nature of the local screening campaign, participants could choose in which ear or ears to perform the test. As a result, 8 participants performed the test sequentially in both ears, hence the gathered dataset consists of 156 observations from 156 ears tested.

The analyzed group of people included both subjects with normal hearing and subjects with varying degrees of hearing impairment, according to WHO standard criteria.
As defined by WHO, the higher the PTA, the higher the hearing impairment. The PTA value to be considered is defined as the mean value, calculated from the measured pure-tone hearing thresholds at the four central frequencies (500, 1000, 2000, and 4000 Hz) tested.
The number of samples for each class of hearing loss is described in Table 1.

WHO criteria	**PTA<25 [dB HL]**	**(1) PTA>25 [dB HL]**	**(2) PTA>40 [dB HL]**	**(3) PTA>60 [dB HL]**	**(4) PTA>80 [dB HL]**
Hearing loss	No impairment	Slight/Mild	Moderate	Severe	Profound
Number of observations / total number of observations	101/156 (64.74%)	37/156 (23.72%)	17/156 (10.9%)	1/156 (0.64%)	0/156 (0%)

Table 1. Sample subdivision according to WHO criteria for hearing impairment. First row: WHO classification in relation to the mean value of the PTA, obtained for the four central frequencies tested (0.5, 1, 2, and 4 kHz) Second row: degrees of hearing impairment. Third row: number of observations fulfilling each criterion, with respect to the total number of observations and relative percentage.

Due to the small number of subjects tested, only one observation corresponds to a person with 'severe' hearing loss, whereas none of the tested subjects have a profound hearing loss.

Audiological screening tests are useful to have an early diagnosis of the problem and are usually proposed to individuals who are not yet aware of having hearing issues or have a problem that is still reduced, while subjects with a severe hearing impairment turn to specialist visits. As a result, the study will focus on creating a classifier that can help distinguishing ‘slight/mild’ or higher degree hearing loss from the results of the speech-in-noise-test to give the patient a first clue about his or her hearing condition. Therefore, only criterion 1 and criterion 2 were addressed.

Each observation has been associated with a target class, according to the selected criterion. In particular, the target variable is 1 (‘fail’) if the criterion is satisfied (PTA> 25 dB HL for criterion 1 and PTA> 40 dB HL for criterion 2) or 0 (‘pass’) otherwise.

3.2 Features extracted

Based on the results gathered during a series of local hearing screening initiatives, a dataset of 156 records and 11 columns has been created.

Each row corresponds to one observation. The first ten columns represent the attributes, i.e., the extracted variables that will be used to predict the new records, while the last column corresponds to the target variable, i.e., the screening result ('pass' or 'fail').

The 10 collected features are here listed and briefly described:

- *Speech reception threshold (‘SRT’).*
  It is the main outcome of the proposed SNT test, as a fairly high correlation with the pure-tone threshold at various frequencies has already been demonstrated (Smoorenburg, 1992) (Jansen, Luts, Dejonckere, van Wieringen, & Wouters, 2013).
- *Raw score of the HHIE-S questionnaire (‘Score’)*, defined as the sum of the points earned for each question. It ranges from 0 to a maximum of 40.
- *Age.*
- *Number of trials (‘#Trials’)*, namely the number of ‘VCV’ proposal before the end of the test.
- *Number of correct responses (‘#Correct’)* obtained during the execution of the test.
- *Percentage of correct responses achieved (‘%Correct’)*, calculated as (#Correct/#Trials)*100.

- *Average reaction time (‘Avg_reaction_time’),* defined as the time (in seconds) between a ‘VCV’ proposal and the selection of a response by the user.
- *Total test duration (‘Total_test_time’*) in seconds.
- *Volume value (‘Volume’),* self-adjustable by the user before starting the test.
- *Division into three classes of hearing impairment, according to the HHIE-S questionnaire raw score (‘Score_classes’)*: class 0 (0 to 8 points, no hearing handicap), class 1 (10 to 24 points, mild/moderate hearing handicap), class 2 (26 to 40 points, significant hearing impairment).

Shapiro-Wilk tests have been performed to verify if the features followed a Gaussian distribution.
As a result, all the variables, except #Trials and #Correct, were not normally distributed ($p<<0.001$).
The defined dataset contains 156 observations and was obtained by integrating data from various sources and cleaning all the lines that contained missing data, for example those cases where the test was not completed or the PTA hearing thresholds were not present for all the central frequencies.

3.3 Correlation Matrix

One of the possible manual ways to explore which variables among those extracted from the test are the best candidates for the prediction of the screening result is to evaluate the correlation matrix. Indeed, variables which are less correlated with the average pure-tone thresholds, could be considered irrelevant as they have no significant impact on the prediction and may be ignored in the classification, whereas the presence of high correlation means that the attribute may be used to predict the target class.
Besides, also correlation between the features themselves may be investigated, in fact if two features present a strong correlation, then they may bring redundant information regarding the predicted class and therefore only one of the two may be considered as a model attribute. The use of this statistical tool indeed may be useful to perform feature selection and reduce the number of attributes involved in the classification task.

The Spearman’s correlation coefficient measures the strength and direction of the relationship between two variables, without requiring any a priori hypothesis about the data

distribution. It ranges from -1 (negative monotonic correlation) to 1 (positive monotonic correlation). As it is a non-parametrical statistical estimate, it may be suitable to describe the association between the considered variables, which appeared to be mostly non-normally distributed, accordingly to the Shapiro-Wilk tests; if this was not the case, the Pearson coefficient could have been used instead.

3.4 Generalized Linear Models

In a subsequent phase, generalized linear models (GLMs) were addressed to establish a mathematical relationship between the attributes and the target variable (i.e., the binary screening result as defined in Section 213.1).

First, the main outcome of the test, i.e. the SRT, was considered as a predictor, then also other attributes were added to improve the goodness-of-fit of the model.

GLMs are a broad family of models for dependent variables with distribution in the natural exponential family, exceeding the limits of linear regression when dealing with non-continuous response variables or response variables with non-constant variance (Agresti, 2013).

GLM class includes, among others, linear regression, Poisson regression and ANOVA models.

A GLM is mainly defined by three blocks (McCullagh & Nelder, 1989):

- The response variable y and its probability distribution (*random component*).
- The explanatory variables ($x_1,\ldots x_k$) and their combination in the model (*systematic component*).
- A function that links the random and the systematic components, explaining the relation between the response variable and the combination of the explanatory variable (*link function*).

Being the dependent variable y binary, i.e. the WHO criteria as related to PTA, the GLMs used during the study are referred to as logistic regression models, characterized by a response variable with binomial distribution and logit function as the canonical link function. This link function returns values between 0 and 1 for arbitrary inputs, therefore it is suitable for binary classification problems.

The goodness-of-fit of the logistic regression model was evaluated by means of a pseudo adjusted R squared, ranging from 0 to 1; while R squared represents the percentage of variance explained by the model, its adjusted version also takes into account the number of independent variables considered, allowing a comparison between models with a different number of predictors and providing information about the power of the model in fitting the regression line.

During the study, the GLMs were used to understand how a variation of a certain amount of a feature could impact the screening result, in order to understand which features were actually the most suitable for the classification task.

3.5 ROC curves

Receiver operating characteristic (ROC) curves were built in order to evaluate the effectiveness in the classification of the presence of hearing loss at various values of speech-reception thresholds, depending on the selected WHO criteria to determine the screening result.

A ROC curve is a two-dimensional plot with 1-specificity (False Positive Rate) on the horizontal axis and the sensitivity (True Positive Rate) on the vertical axis, for all possible cut-off values, that allows to express the information content of confusion matrixes and the performance of classifiers in a visual way. In the present thesis work, the speech reception threshold was varied from 10 to -20 dB.

The point (0,1) represents the ideal classifier, the point (0,0) represents a classifier which predicts class 0 (‘no hearing loss’) for all the records, whereas point (1,1) corresponds to a classifier predicting 1 (‘hearing loss’) for all the observations. Hence, the SRT cutoff value corresponding to the best discrimination is the one closest to the point representing the ideal classifier; examined individuals with SRT higher than the cut-off fail the screening test (presence of hearing loss) whereas subjects with SRT below the cut-off value successfully pass the test.

The area beneath the ROC curve (AUC) is a suitable parameter to be used to compare classifiers as it is a measure of the capability to distinguish between classes. Indeed, the greater the AUC, the better the classification performance.

3.6 Classification

This study investigates the capability of the developed speech-in-noise test to discriminate patients with 'slight/mild' or higher degree of hearing loss using machine learning algorithms, with the ultimate goal of implementing a smartphone application for adult hearing screening.

The objective is therefore to classify observations in 'pass' or 'fail' using a machine learning approach based on the variables extracted by the speech-in-noise test software to predict the result of the PTA in terms of level of hearing impairment. Since a screening test is a preliminary assessment useful to make the subject aware of the problem before it leads to serious consequences, the WHO criterion based on an average hearing threshold of 25 dB HL for the identification of 'slight/mild' to more severe hearing loss (criterion 1) has been investigated.

Several classification algorithms have been investigated: Decision Trees (DTs), Support Vector Machines (SVMs), Logistic regression, K-Nearest-Neighbors, ensemble Logistic regression, Random forests and Gradient boosting. Their peculiar characteristics are described in the next sections of this chapter.

Each classifier has been evaluated by feeding it with different combinations of the available features that will be discussed in the next chapters.

3.6.1 Data Preparation

In order to proceed with a multivariate analysis of the test classification performance, with the aim of limiting overfitting, the dataset was randomly split into a training and a test partition, with a ratio of 80% (124 ears) and 20% (32 ears) respectively. Stratification was used to maintain the proportion of "pass" and "fail" present in the original dataset also in the training and test partitions.

The dataset underwent a scaling procedure to ensure that the attributes of the model satisfied the requirement of certain learning algorithms that exploit the concept of distance or similarity (i.e., k-Nearest-Neighbors and Support Vector Machines), to have Gaussian distributed individual features (e.g. zero-mean and unit-variance).

A standard scaling procedure (Z-score Normalization) was chosen to obtain scaled features $x_{ij}{}'$ with zero-mean and unitary variance, defined as:

$$x_{ij}{}' = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$$

Where $\bar{\mu}_j$ and $\bar{\sigma}_j$ are the sample mean and the sample standard deviation of each attribute $\boldsymbol{a_j}$, i.e. the features considered (Vercellis, 2009).

To allow model optimization and avoid overfitting, a validation phase was needed. To ensure that the results obtained do not depend on the choice of train and validation partition, especially when dealing with small datasets, a k-fold cross-validation has been introduced. Cross validation requires k iterations and consists of partitioning the dataset into k disjoints subsets $S_1, \dots, S_k$. The number of partitions k was set to 5 during the study, so that each subset was large enough to be representative. At the i-th iteration, the subset S_i is the validation set whereas the training set is made by the union of all the other subsets ($T_i = \bigcup_{m \neq i} S_m$). This procedure makes sure that each single observation appears k-1 times in the training set and only once in the validation set. The classifier is applied k times. At each iteration the model is fitted on the training set and the accuracy is evaluated on the corresponding validation set, as schematized in Figure 6. Finally, the accuracy is computed as the mean of the single accuracies over the k iterations. Stratification has been used to maintain the same proportion of observations for each target class in each partition. After fitting, each model was tested on the test set.

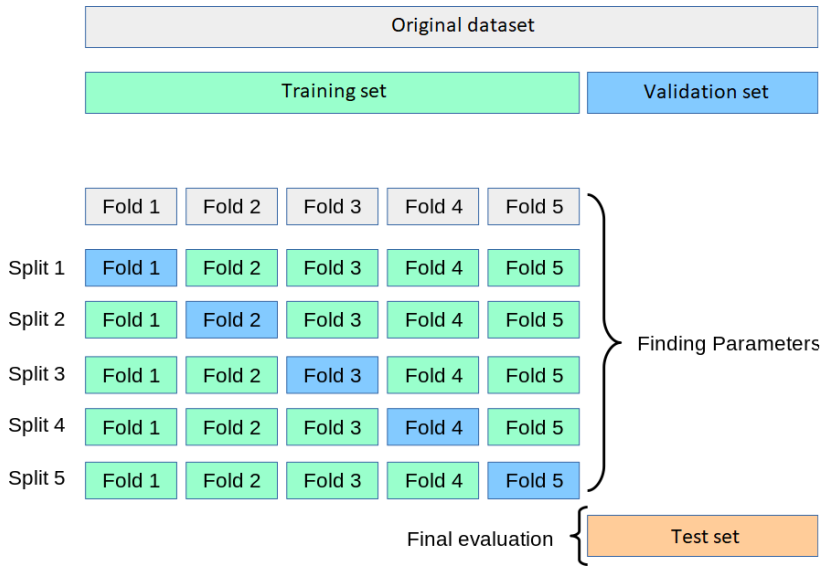


Figure 6. K-fold cross-validation procedure. The training set is partitioned into k subset. At each iteration, k-1 folds are used as training set and the remaining one is used as validation set.

3.6.2 Decision Trees

Decision trees (DTs) are supervised machine learning algorithms that can be applied to both classification and regression issues. Classification trees are widely used, able to work for both categorical and continuous variables, fast, robust with respect to outliers and most of all, they generate interpretable separation rules to separate observations belonging to different classes. A tree is generated during the training phase, following a top-down induction scheme. All the observations of the training set are initially contained in the root node then, following an iterative heuristic process, they are split into different disjoint subsets (branching) according to some separating rules based on attribute values, until a stop criterion is fulfilled.

Considering univariate trees, also called axis-parallel trees, each splitting rule can be defined by the formula $X_j \leq b$ or $X_j \geq b$, hence it is based on a condition that takes into account only one attribute X_j. Focusing on binary trees, each splitting rule divide the observations in the parent node into two subsets, creating two branches, according to a threshold value set on a given attribute.

Each node is associated with a heterogeneity index I(q), also called impurity index, which is a measure evaluated in order to select a splitting criterion, so as to divide the data in the best way. The impurity index is defined in order to reach its minimum when all the observations in the node belong to the same class, whereas it reaches its maximum when the observations belonging to the node are evenly distributed among the target classes (i.e., all the classes in the considered node have the same probability). The tree heterogeneity index commonly used are the misclassification index, the entropy index and the Gini index.

In particular, the Gini index of a node q is defined as:

$$Gini(q) = 1 - \sum_i p_i^2$$

Where p_i is the probability associated with class i.

Hence, the index is 0 if the node includes samples belonging to one class only (pure node) and 0.5 if each class in the node is equiprobable.

The selection of the best splitting rule is based on a greedy algorithm that is the minimization of the impurity of the nodes with respect to the target variable, going down the tree.

Indeed, the best separation rule is the one that provides the smaller Gini index, defined as the weighted sum of the impurity value of each partition.
Considering a binary tree, the Gini index for a separation rule based on attribute A is:

$$Gini_A(P) = \frac{|P_1|}{|P|} * Gini(P_1) + \frac{|P_2|}{|P|} * Gini(P_2)$$

Where P is the parent node, while P_1 and P_2 are the two child nodes generated by the separating rule.
When making a prediction, the new record follows a path down the tree, according to the sequence of splitting rules, which ends with a leaf node. Each bottom node, is associated with a particular class ('pass' or 'fail'), following a majority voting procedure. Hence, the new observation is associated to the class belonging to the majority of the observations of the training set that have fallen into the node.
The prevention of overfitting is one of the key problems to be addressed when dealing with tree-based algorithms. Therefore, some constraints can be set on tree parameters, like the minimum number of samples required in a leaf node, the maximum number of leaf nodes or the maximum vertical depth of the tree. An alternative to add constraints could be to advance a pruning procedure i.e. to avoid excessive growth of a tree (pre-pruning) or to remove nodes involving rules with lower importance after the tree has grown (post-pruning).
DTs with no growth limit and with different levels of maximum achievable depth (3, 4 ,5) have been analyzed in the following chapter.

3.6.3 Support Vector Machines

Support Vector Machines (SVMs) are supervised machine learning algorithms that can be used to solve both classification and regression problems and provide good performance even with small datasets. The main goal of SVM algorithm is to find a hyperplane in a n-dimensional space, where n is the number of attributes in the dataset, able to classify the observations in the best way. If the problem deals with only two features, then the hyperplane is simply a line. In this study, instead, having to deal with 10 features or a subset of them, hyperplanes are much more complex and difficult to represent.
The best hyperplane, leading to the optimal discrimination between classes, is the one that guarantees the maximum margin between the nearest data points and the plane itself, hence,

better segregates the two classes. The position and the orientation of the plane is influenced from the observations that are closer, also called support vectors. (Mohri, Rostamizadeh, & Talwalkar, 2018)

The hyperplane acts as a decision boundary: observations falling on one side will be classified as 0 whereas observations on the other side will be classified as 1. If the data are non-linearly separable, a SVM can map the space into higher dimensions by means of kernel functions, so that a non-linear separation in the original features space corresponds to a linear separation in the transformed space (with higher dimensionality).

More and more complex kernels can be derived by linear combination of simple ones.

Example of decision surfaces obtained using different type of kernels for a SVM considering only the features 'SRT' and 'Age' are reported in Figure 7. This particular example shows that the classifier implementing a polynomial kernel is better in discriminating the 'blue' class (only one misclassification), however the error in classifying points belonging to the 'red' class (i.e., individuals with hearing impairment) is higher compared to the linear kernel, meaning that a lot of ill subjects are classified as healthy.

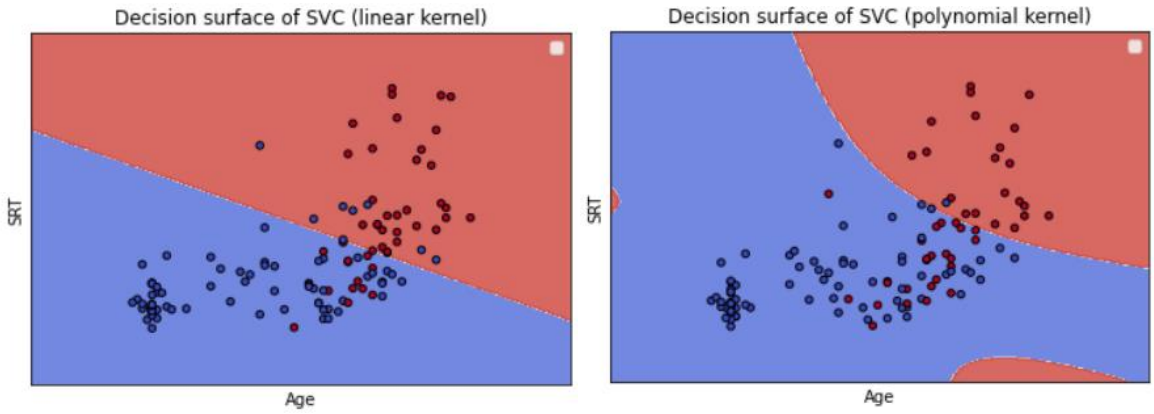


Figure 7. Decision surfaces for Support Vector Machine classifiers implementing different kind of kernels.

Feature scaling is required when dealing with SVM to avoid that attributes with larger range dominates over the others (i.e., 'Avg_reaction_time') and simplify kernels calculation.

3.6.4 Logistic Regression

Logistic Regression is a machine learning approach that allows to solve binary classification problems transforming them into linear regression ones. Linear regression cannot simply be

used in binary classification problems as its output values are unbounded, whereas the desired output must be either 0 or 1.

Considering a classification problem involving the target variable $y \in [0,1]$, as in the screening procedure here considered, and the features vector $\boldsymbol{x} = [x1, x2, \ldots xn]$, a logistic regression technique is based on the concept of conditioned probability $P(y|\boldsymbol{x})$.

Predicted values are mapped to be in the range 0-1 by means of a canonical function called logistic function, therefore predicted values may be interpreted as probability classes.

The logistic function, or sigmoid function, is a S-shaped curve that can map a real number into a range of values comprised between 0 and 1. Its shape is reported in Figure 8.

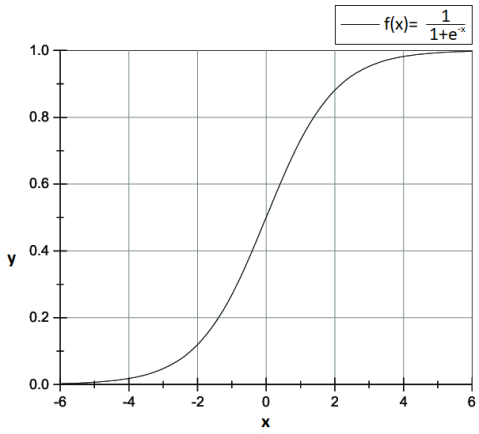


Figure 8. Standard logistic function curve, presenting a s-shape, ranging from 0 to 1.

Indeed, the canonical link function for logistic regression is the logit function, defined as:

$$logit(p) = log\left(\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}\right) = \beta + \boldsymbol{xw}$$

Linearly depending on the explanatory variables.

To predict the class of a new record, the estimated probability must be considered. If the probability is higher than a threshold value (e.g. 0.5) the observation is classified as 1 ('fail'), otherwise as 0 ('pass').

3.6.5 K-Nearest-Neighbor

K-Nearest-Neighbor (KNN) is a supervised learning algorithm based on the concept of distance between data points, that can be used both in regression and classification problems.

Indeed, it assumes that observations belonging to the same class are close to each other. It is a non-parametric tool because no a priori hypothesis on the data distribution is required. The prediction of a new record using KNN requires the following steps:

I. Choice of the number of neighbors k to consider. k was set to 5 in the thesis work.

II. Calculation of the distance between the new observation and each data point in the features space. The distance can be calculated by means of different metrics, i.e. Euclidean, Hamming, Minkowsky or Manhattan.
In particular, the Euclidean distance between two points $p_1(x_1, y_1)$ and $p_2(x_2, y_2)$, which is the one chosen throughout the study, is defined by the formula:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

III. The computed distances are sorted from the smaller one to the bigger one.

IV. Only the k data points with the lower distances are considered. The new observation will be assigned to the class to which belongs the majority of the k nearest data points (i.e., the mode of the k classes). In regression problems instead, the observations will be associated to the average of the labels of the k data points.

An example of the concept of majority voting in KNN technique is described in Figure 9.

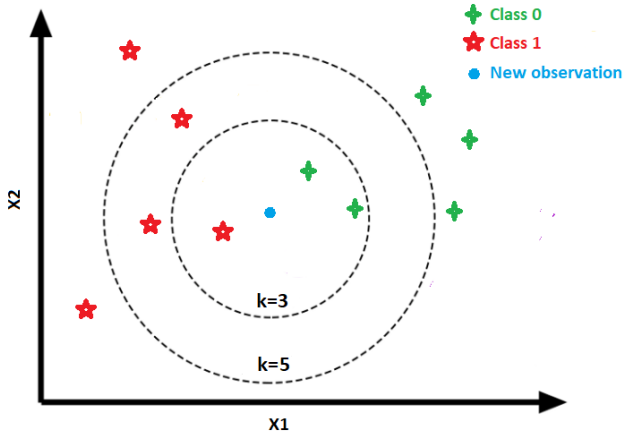


Figure 9. KNN classification example showing the influence of the number of neighbors k in the prediction output. Considering k=3 the new observation will be associated to class 0 whereas considering k=5 the observation will be associated to class 1.

3.6.6 Ensemble Logistic Regression

Ensemble techniques are machine learning techniques that combine the decisions from multiple models (base models or 'weak learners') to obtain better classification performance with respect to a single model, trying to reduce its bias and/or variance.
The three main methods for ensemble learning include stacking, bagging and boosting.

Stacking techniques apply the idea to train several different weak learners, therefore they are referred to as heterogeneous techniques, and use the predictions returned by these models to train another model. Stacking techniques are based on two levels of processing.
Models in the 0-level are typically different and fitted on the same training set in parallel (independently). The approach used during the study involved a 'k-fold cross-training', indeed the training set has been split in k partitions and at each iteration the base model is trained on k-1 fold and tested on the remaining fold, so that in the end predictions for each observation of the training set have been obtained and all the predicted probabilities were used to train the meta-classifier.
Then, the meta-classifier is trained on the predicted probabilities that are outputs of the base models and therefore learns to make predictions on new data based on the multiple predictions returned by the 0-level models.
Figure 10 shows a summary scheme of the stacking algorithm employed in the thesis.
A number of fold equal to 5 was chosen, then the four machine learning algorithms previously discussed, namely a decision tree, a support vector machine, a logistic regressor and a k-nearest neighbor classifier, were selected as base models.
Besides, another logistic regressor was used as meta-classifier to learn how to combine predictions of the previous models, by using as input the probability values from base-models, and then to predict new data. The final prediction performances were evaluated using the test set.

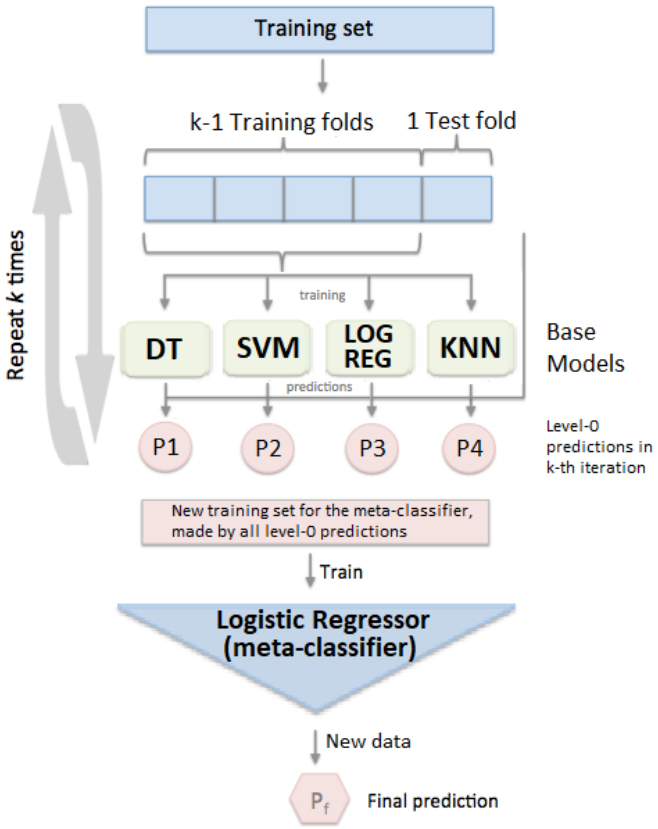


Figure 10. Summary scheme of an ensemble logistic regressor implementing stacking approach.

3.6.7 Randoms Forests

Random forest is an ensemble machine learning algorithm based on trees, following a bagging approach.

Bagging techniques consist in fitting several learners of the same nature (hence are homogeneous techniques) in parallel and average their predictions in order to obtain a result with reduced variance with respect to the single classifier. Each single base model can work on almost independent datasets by means of bootstrapping, which is a statistical technique that, starting from the initial dataset, generates a certain number of subsets, selecting observations with replacement.

In Random forests, each individual DT works in parallel on a random subset of the original data set and the final output of the model is the combination of the outputs of single decision trees by means of majority voting, which means that the observation is associated to the most popular class. A simple schematic is reported in Figure 11.

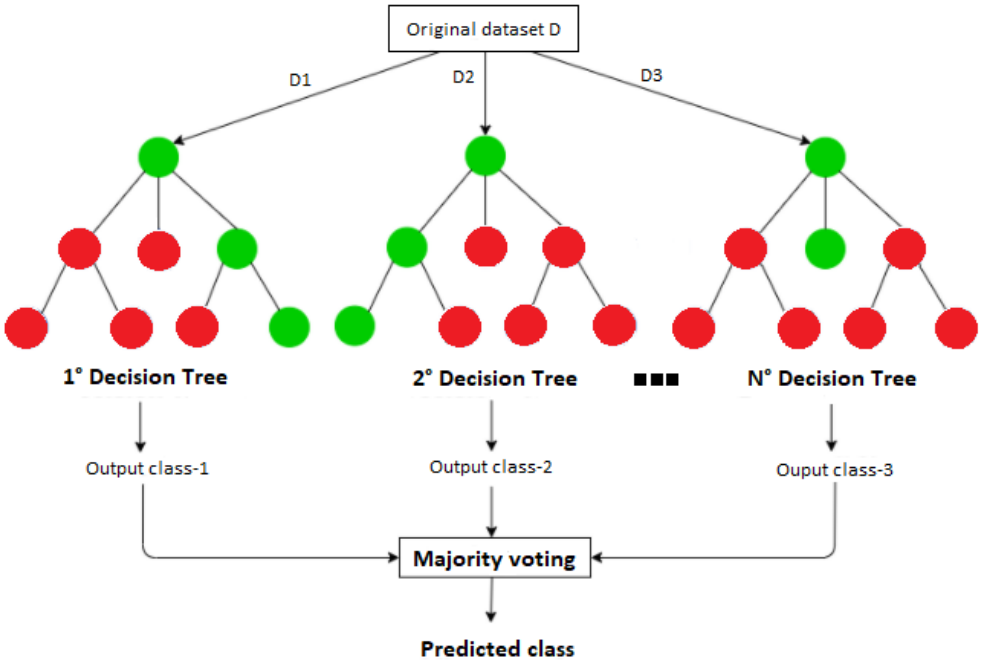


Figure 11. Random forest scheme.

When dealing with forests it is easy to measure the importance of a feature, indeed the relative score for each feature can be computed, so that features with the lower score i.e. less contributing to the classification, can be removed. Important parameters that can be tuned to optimize the model are for example the number of trees in the forest and the number of features to be used in a splitting rule.

Increasing the size of the forest increase the stability, because the decisions of a higher number of models are taken into account, however it also increases the time required for computation. During the analysis, Random forests made by 10, 50 and 100 trees were considered.

3.6.8 Gradient Boosting

Gradient boosting is the last ensemble machine learning technique considered, exploiting the boosting approach to combine a certain number of classifiers to form a stronger learner. In boosting techniques, which are sequential methods, weak models are not independent anymore but the training of the model at a certain iterative step depends on the models previously fitted. In particular, each new model added gives more attention to the observations that were misclassified by the previous models, so that, in the end we obtain a

classification algorithm with lower bias. It can be used to solve both regression and classification problems.

Gradient boosting commonly exploits a tree-based additive approach; each new tree sequentially added to the model corrects and improves the classification performance of the precedent model. Following the hypothesis of boosting, starting from a single weak classifier (e.g. Decision tree), new learners (e.g. new trees) are iteratively added over time in a sequential way and therefore fitted on a modified version of the training set.

In AdaBoost, which is one of the first boosting technique ever implemented, after the evaluation of a DT, weights are associated to the observations according to the difficulty of classification. Observations that are hardly classified are associated to higher weights with respect to observations that are more simply classified. Then, the following tree added is grown on this new weighted training set, with the aim to improve the predictions of the previous tree.

Gradient boosting, which is a generalization of AdaBoost, instead focuses on the minimization of a customizable loss function, like the gradient descent method implemented in neural networks. Indeed, for each tree the loss function is calculated, then, a gradient descent is performed to reduce that function and parameters of the learner are modified accordingly. Results are appended to the previous ones. The iteration is repeated until either a maximum number of learners is reached, or the loss function is minimized.

There are different ways to improve classification and try to avoid overfitting phenomenon, such as to impose limits on the depth of the single tree or choosing the number of trees and learning rate.

At each iteration, trees can be fitted on random partitions of the training set, in that case the algorithm is defined as stochastic gradient boosting (Friedman, 2002).

Both classical gradient boosting and stochastic gradient boosting together with shrinkage (i.e., the use of small learning rates) have been investigated and will be discussed in the next chapters.

3.6.9 Metrics to evaluate the performance of a classifier

The performance of each classification model under study was evaluated by measuring the following parameters:

- *Accuracy on the training set and accuracy on the test set*, describing the number of correct assessments (either 'pass' or 'fail') in the portion of the dataset considered.

$$Acc = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ observations}$$

- *Area under the curve (AUC):* measuring the ability of a classifier to discriminate between the 2 classes; the higher the AUC, the better the model performance, indeed, the perfect test yield AUC=1, meaning that all points are correctly classified, whereas a test with 100 % misclassifications has AUC=0. AUC is a useful metric when dealing with unbalanced classes.
- *True Positives (TP):* number of positive observations ('fail') correctly predicted.
- *False negatives (FN):* number of positive observations ('fail') erroneously predicted (type II error).
- *Specificity:* number of observations defined as 'pass', that were truly 'pass'.

$$Spec = \frac{TN}{TN + FP}$$

- *Sensitivity (or Recall or True Positive Rate (TPR)):* 'fail' samples correctly detected with respect to all the 'fail' observations in the dataset, defined as:

$$TPR = \frac{TP}{TP + FN}$$

- *FNR (False Negative Rate)*, defined as:

$$FNR = \frac{FN}{TP + FN}$$

- *Precision:* proportion of correctly classified 'fail' samples, i.e. the accuracy of the positive prediction, defined as:

$$p = \frac{TP}{TP + FP}$$

Hence, precision indicates the records labeled as having hearing loss that were hearing impaired indeed. Precision alone is not enough because either having one single 'fail' observation or FP=0 would result in p=100%.

– *F-measure*, defined as the harmonic mean of precision and sensitivity, giving a combined idea about the two metrics:

$$F = \frac{(\beta^2 - 1) * TP * p}{\beta^2 * p + TP}$$

With β factor regulating the relative importance of precision with respect to TP. Indeed, F-score is maximum when precision equals sensitivity.

The confusion matrixes were built by using the probability closest to the point (0,1) in the calculated ROC curve as a cut-off for the prediction probabilities, instead of default value of 0.5. Indeed, a record having an output probability higher than the threshold is classified as 'fail', whereas a record with an output probability lower than the threshold is classified as 'pass'.

In the design of an application for medical screening the main concern is to correctly classify ill people, therefore the number of false negatives (i.e., people classified as not hearing impaired but actually suffering from hearing loss) must be reduced. Besides, sensitivity is another useful metric, showing how good the test is in detecting hearing impairment. However, a correct classification of people with no impairment is also important, hence specificity is important too.

Finally, the generalization properties of the model must be considered, in fact, the model has to learn a relation from the training set, but it also has to be able to predict the class (i.e., the screening results) of new records, trying to minimize errors. Thus, the classifier must be flexible, and overfitting must be avoided. Therefore, accuracies on both training and test sets need to be considered.

Figure 12 shows an example of overfitting; the model is able to fit very well the training set (past data), however it lacks flexibility, making errors in the prediction of new data. A good model should instead reach the best trade-off between training and prediction errors.

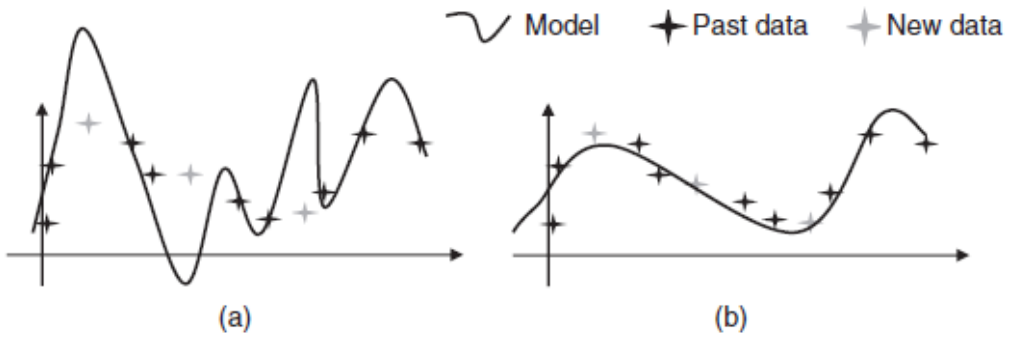


Figure 12. Panel a represents an example of overfitting: the model is very good at fitting past data (accuracy on the training set is 100%) but it has very poor generalization capabilities (great error on the test set, i.e. low accuracy on the test set). Panel b shows instead a good model that is able to guarantee a trade-off between explanation of past data and generalization of future data (Vercellis, 2009).

A clue to verify the presence of overfitting is to check the difference between the accuracies of training set and test set.

The initial splitting of the dataset into training and test partitions can have a huge impact on the model performance when dealing with a small number of observations, indeed the model are data-driven. In order to contrast this issue, 1000 iterations of the model optimization process have been performed both by changing the initial random subdivision into training set and test sets (keeping the partitions 80%-20%) and the 5-fold cross-validation subsets. Therefore, for each indicator of the model performance, the average value and the standard deviation over 1000 iterations have been calculated.

4. Results

4.1 Statistical characterization of test variables

In order to better characterize the composition of the sample under investigation, the distribution of the variables extracted by the speech-in-noise test software have been analyzed according to age, gender of the participants and ear tested.

The dataset has been divided in two groups according to sex (male or female), ear tested (right or left) and in three groups for what concerns the age (Young, Adults, Elderly).

Being the variables considered not normally distributed, a two-sided Wilcoxon rank sum test was computed for each subdivision in order to check the equality of the medians for the two independent unequal-sized samples. Significance level was set to 5% during all the analysis.

4.1.1 Statistical characterization of test variables: left vs right ear

Due to the nature of the experimental campaign, participants could choose in which ear(s) to perform the test. Only 8 participants decided to perform the test sequentially in both ears whereas the others performed the test only in one ear. As a result, the dataset is composed by 88 samples related to the test executed on the right ear and 68 samples to the left ear.

Table 2 illustrates the distribution of the dataset observations according to the ear tested. Only the average reaction time reveals a statistical difference between right and left ear ($p<0.05$), while all the other features distributions show no significant difference.

	Right ear	**Left ear**	**p-value**
SRT [dB SNR]	-12.86 ± 7.02	-9.15 ± 9.73	0.07
Mean PTA [dB HL]	20 ± 14.61	19.37 ± 17.26	0.83
# Trials	80 ± 14.45	74 ± 18.24	0.16
# Correct Answers	72 ± 14.08	66 ± 17.82	0.19
%Correct Answers	90.19 ± 3.56	89.89 ± 4.33	0.60
Total_test_time [s]	221 ± 53.31	241.5 ± 65.26	0.26
Avg_reaction_time [s]	1.63 ± 0.66	1.93 ± 0.93	0.02
Volume	0.5 ± 0.10	0.5 ± 0.12	0.15

Table 2. Distribution of the dataset observations according to the ear tested. The column on the far right shows the p-value of Wilcoxon rank sum test between the features' populations related to the right ears and the one related to the left ears.

4.1.2 Statistical characterization of test variables: male vs. female

The subdivision of the sample by gender is fairly disproportionate, indeed the number of women (102) who took part in the experimental screening campaign is more than double the number of men (46). Despite so, none of the features considered exhibit a significant difference in the distribution due to the gender. Results can be appreciated in Table 3.

	Male	**Female**	**p-value**
SRT	-11.16 ± 7.19	-12.40 ± 8.98	0.75
Mean PTA	20 ± 15.38	20 ± 16.77	0.59
Age	60 ± 20.77	58.5 ± 20.21	0.91
Score	4 ± 7.96	3 ± 8.18	0.75
# Trials	79 ± 14.80	76.5 ± 16.94	0.45
# Correct Answers	70.5 ± 14.16	69 ± 16.62	0.39
%Correct Answers	90.28 ± 2.32	90 ± 4.42	0.43
Total_test_time [s]	231.5 ± 43.65	235.5 ± 64.58	0.62
Avg_reaction_time [s]	1.65 ± 0.54	1.82 ± 0.88	0.17
Volume	0.5 ± 0.068	0.5 ± 0.13	0.37

Table 3. Distribution of the dataset observations according to gender of the tested person. The column on the far right shows the p-value of Wilcoxon rank sum test between the features' populations related to male and the one related to female subjects.

4.1.3 Statistical characterization of test variables as a function of age

Age of the subjects in the gathered sample covers a very wide range (20-89 years old) therefore, it was decided to divide the observation cases into three different groups, according to the age of the participant:

- *Young:* 20<Age≤25 years old ($N_{subjects}$=36; N_{ears}=37; mean=23.78; std=1.12)
- *Adults:* 25<Age<60 years old ($N_{subjects}$=46; N_{ears}=46; mean=47.35; std=9.90)
- *Elderly:* Age≥60 years old ($N_{subjects}$=66; N_{ears}=73; mean=70.90; std=7.25)

The choice of this subdivision was made after consulting other age-based subdivisions present in the literature (Heidari, Moossavi, Yadegari, Bakhshi, & Ahadi, 2018) (Paglialonga, Tognola, & Grandori, 2014).

As it can be noticed from Table 4, the partition by age is quite well defined, with almost all features distributions being statistical different considering different age groups.

	Young	Adults	Elderly	p-value		
				(a)	(b)	(c)
SRT	-15.86±2	-12.86±5.39	-6±9.08	9.87e-05	2.33e-07	1.87e-13
Mean PTA	0±5.67	20±8.65	31.25±12.39	4.81e-13	4.29e-08	5.66e-17
Score	0±2.58	4±5.83	4±9.87	9.58e-05	0.21	1.56e-06
# Trials	86±13.01	79±11.28	72±16.72	3.49e-04	0.002	3.47e-08
# Correct Answers	79±11.88	72.5±10.86	63±16.14	1.64e-04	5.31e-04	3.82e-09
%Correct Answers	91.40±1.05	90.53±1.59	88.71±4.9	0.006	3.92e-07	1.40e-11
Total_test_time [s]	213±41.47	229±49.3	254±69.26	0.32	0.11	0.02
Avg_reaction_time [s]	1.33±0.21	1.6±0.51	2.27±0.86	2.11e-04	8.44e-07	1.06e-13
Volume	0.5±0.11	0.5±0.9	0.5±0.12	0.0024	0.96	6.99e-04

Table 4. Distribution of the dataset observations according to age of the tested person. The column on the far right shows the p-value of Wilcoxon rank sum test between Young ad Adults (a), Adults and Elderly (b) and Young and Elderly (c).

The only cases not showing any statistical difference ($p > 0.05$) concerned the score and the volume between Adults and Elderly and finally the total test duration between the groups Young and Adults and the groups Adults and Elderly.

SRT, 'meanPTA', 'Score' and 'Avg_reaction_time' increase as age increases. On the other hand, the number of trials, the number of correct responses and consequently the percentage of correct responses diminish as age increases. As an example, the distribution of the number of correct responses for the three groups is shown in Figure 13.

It can be clearly seen that, as previously stated, the values of this variable decrease significantly when we consider adult subjects with respect to young subjects and further decrease when we consider elderly patients.

Volume, instead, appears to be slightly independent from the age parameter, with a value distributed around 0.5, which is the default value.

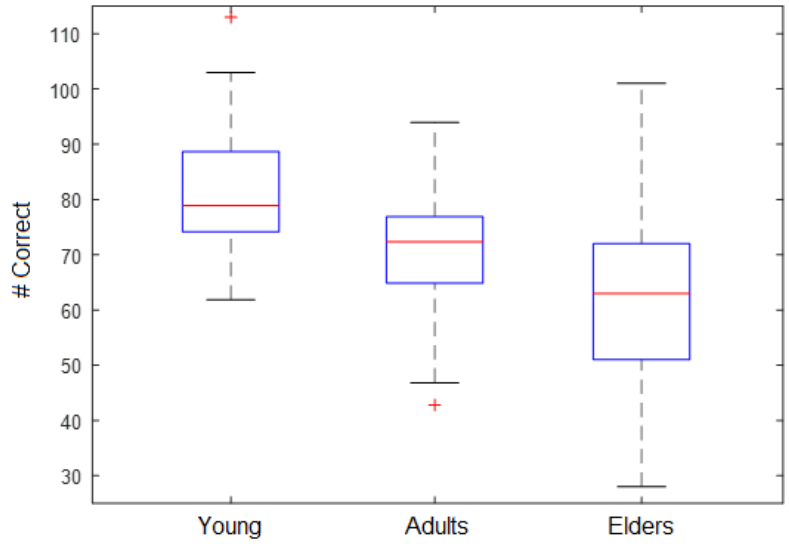


Figure 13. Distributions of the number of correct responses considering a partition of the dataset into three groups, by age.

4.2 Correlation between test variables (features) and outcome statistical feature characterization

To further investigate the presence of associations between variables in the dataset and try to summarize the data, the correlation matrix for the variables identified during data set collection was calculated. The computed matrix is displayed in Figure 14.

Each cell (i,j) of the matrix represents the Spearman's correlation coefficient between variable Xi and variable Xj, ranging from r=1 (strong positive correlation) to r=-1 (strong negative correlation).

The Spearman's coefficient is the most suitable coefficient as it does not hold any assumption about the data distribution and therefore it can be used to evaluate correlation between variables that are not normally distributed.

A variable is always fully correlated to itself, as a consequence, the diagonal entries of the matrix are equal to 1. Since the correlation matrix is symmetrical, in order to simplify its visualization, only the lower half has been shown.

Alongside with the correlation coefficient, also the matrix of p-values has been calculated. If an element outside the diagonal of the correlation matrix has a p-value lower than the significance level alpha (set to 0.05), then the corresponding correlation must be considered significant.

Among the features analyzed, the volume resulted to be the less correlated one, with a significant dependency ($p<<0.01$) only with the average reaction time referred to a weak negative correlation, whereas the SRT was significantly correlated ($p<<0.001$) with almost all the other variables except the total test duration and the volume.

Clearly, for how the test was implemented, the number of trials and the number of correct responses showed a very high correlation (Spearman's coefficient = 0.99), as more correct responses lead to an higher number of trials to reach 12 reversals, and therefore the end the test.

As the purpose of the study is to use the output variables of the speech-in-noise test to predict the result of the PTA, it can be meaningful to focus on the column related to the variable 'meanPTA'. As it could be expected, SRT, which is the primary outcome of the developed test, turned out to be one of the variables most correlated (Spearman's coefficient = 0.66) with PTA. This reinforces the concept that the SRT can be used as a discriminant variable for an audiological screening test.

Together with it, also 'Age' and 'Avg_reaction_time' presented a quite high and positive correlation with the PTA, whereas '#Trials', '#Correct' and '%Correct' are correlated too, but in a negative way.

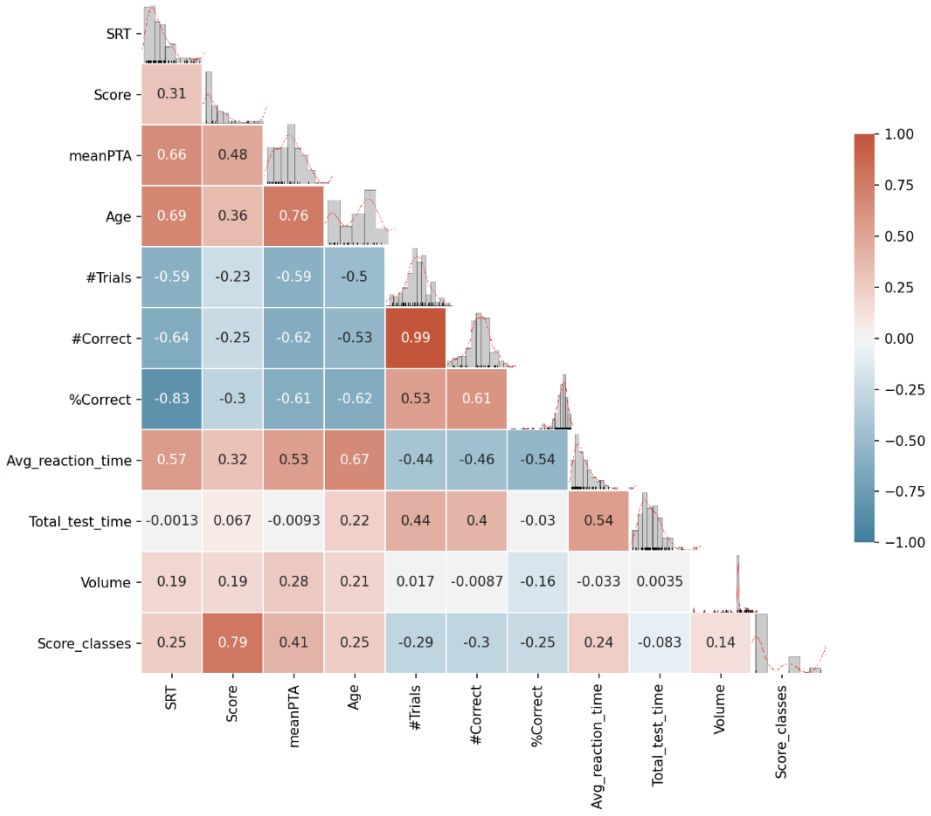


Figure 14. Correlation matrix between the features extracted from the dataset. The corresponding correlation coefficient is reported in each cell, while the strength of the relationships is represented by the color of the cell. Cells colored in red are positively correlated whereas blue cells represent features negatively correlated.

To address any possible relationship between the test output variables and their distribution in 'pass'/ 'fail' according to the two WHO criteria, the scatter plots for each combination of paired features, as well as the distribution of each feature in the two classes, have been computed.

Results are shown in Figure 15 (criterion 1) and Figure 16 (criterion 2) respectively.

On the main diagonal of the matrix, the histogram with the distinction in the two classes of each single feature extracted from the test results is reported. This representation demonstrates that some variables are more suitable to classify the tested cases according to the WHO criteria, with respect to the others. In particular, the feature that can be considered better for classification purposes is, as expected, the SRT, as it presented two quite well distinguished distributions in the two classes.

In addition to it, also age, the average reaction time, the number of trials, the number of correct responses and as a consequence, even the derived percentage of correct responses,

can be considered suitable features for the classification task, because their distribution presented a quite defined distinction between the two classes. Indeed, focusing for example on the scatterplot related to SRT combined with age, whose magnified version is reported in Figure 17, we can notice that data points are relatively grouped into the two classes (‘pass’/ ‘fail’), generating distinguishable clusters of data. In fact, most of the ‘fail’ records can be identified at high values for both SRT and age. Another example is the scatterplot of the number of correct responses and the average reaction time, reported in Figure 18, showing a quite distinguishable cluster of ‘fail’ points for low values of correct responses and high values of average reaction time.

On the other hand, scatterplots related to features like total test duration and volume, displayed in Figure 19, do not present such sharp distinctions between classes and therefore are less useful for classification purposes, as the distribution of ‘fail’ is wider and therefore cannot be distinguished from the ‘pass’ distribution.

The logic can be applied for both criteria; although for criterion 2 there are fewer ‘fail’ records, a sharper division into the two classes can be also noted for the same variables highlighted for criterion 1.

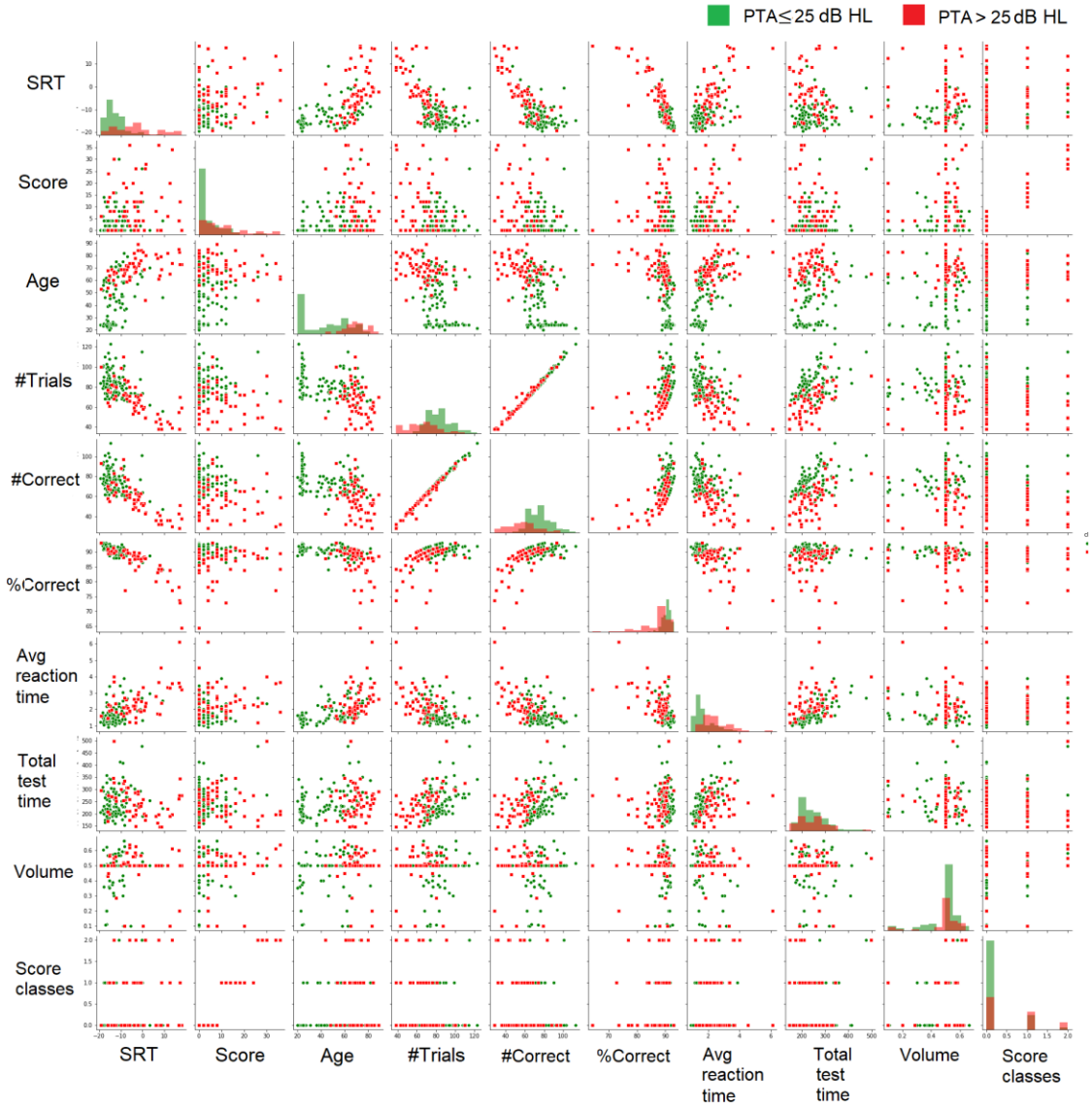


Figure 15. Scatter plots of paired features referred to criterion 1. Along the main diagonal the distribution of each single feature in the two classes is represented. Green marks: tested ears with PTA ≤ 25 dB HL. Red marks: tested ears with PTA > 25 dB HL.

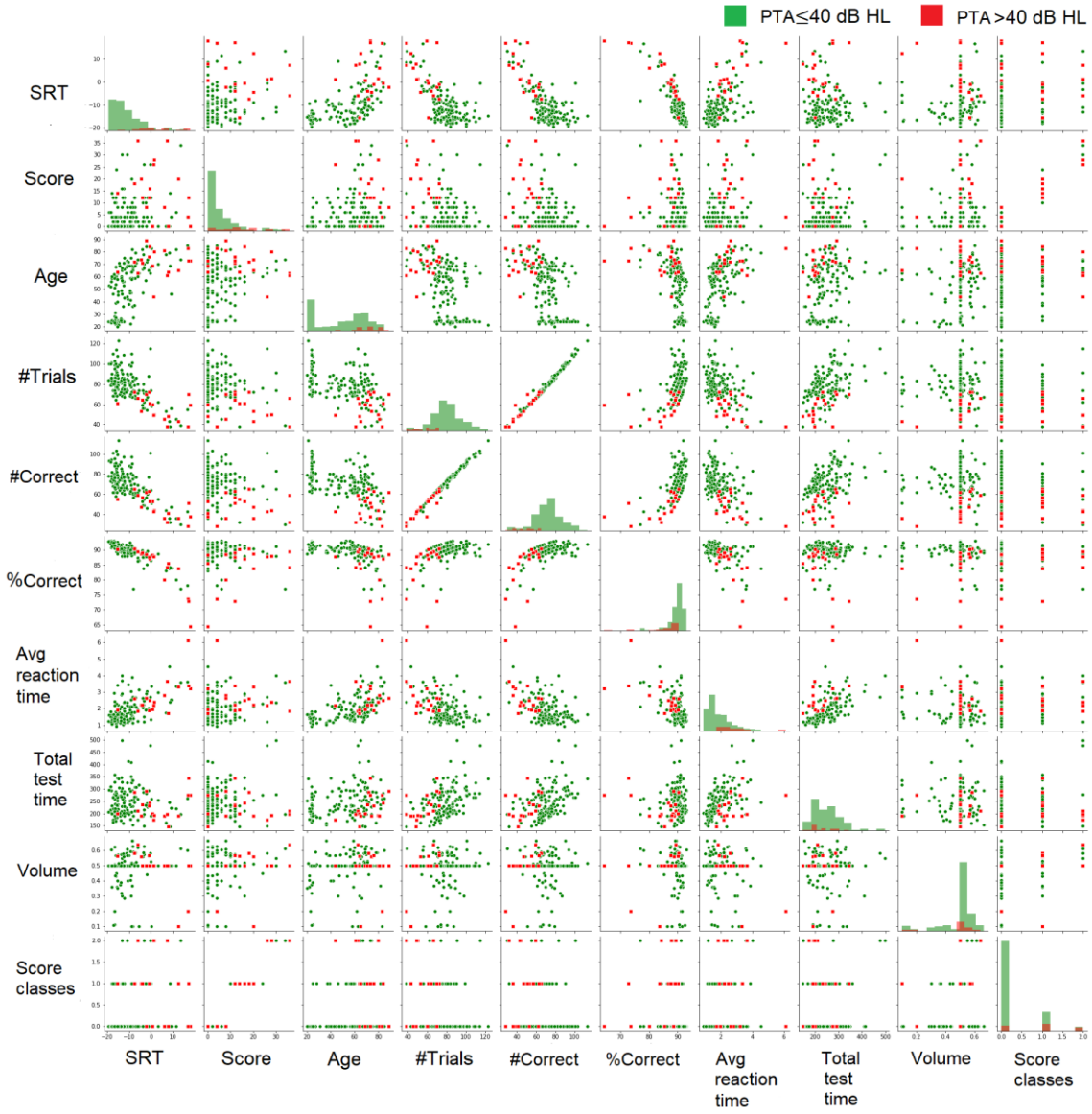


Figure 16. Scatter plots of paired features referred to criterion 2. Along the main diagonal the distribution of each single feature in the two classes is represented. Green marks: tested ears with PTA ≤ 40 dB HL. Red marks: tested ears with PTA > 40 dB HL.

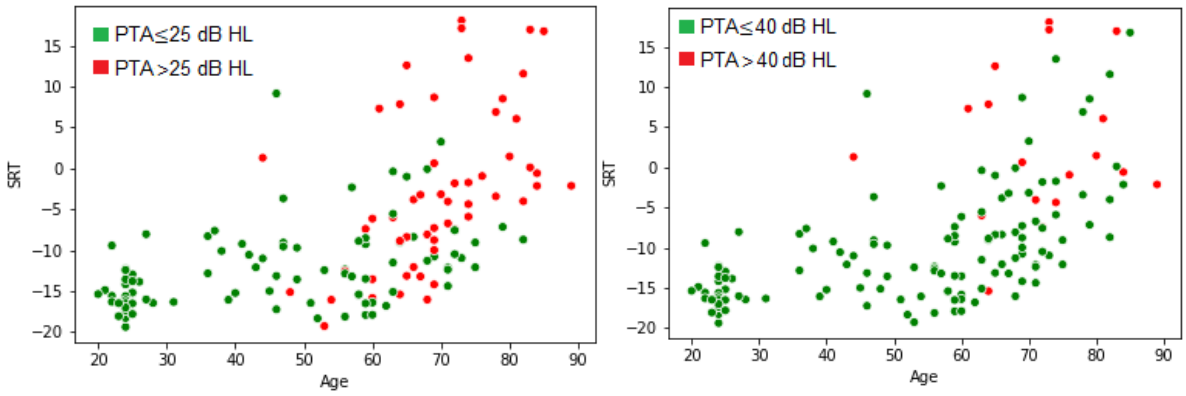


Figure 17. Scatterplots of the SRT combined with Age for criterion 1 (left panel) and criterion 2 (right panel).

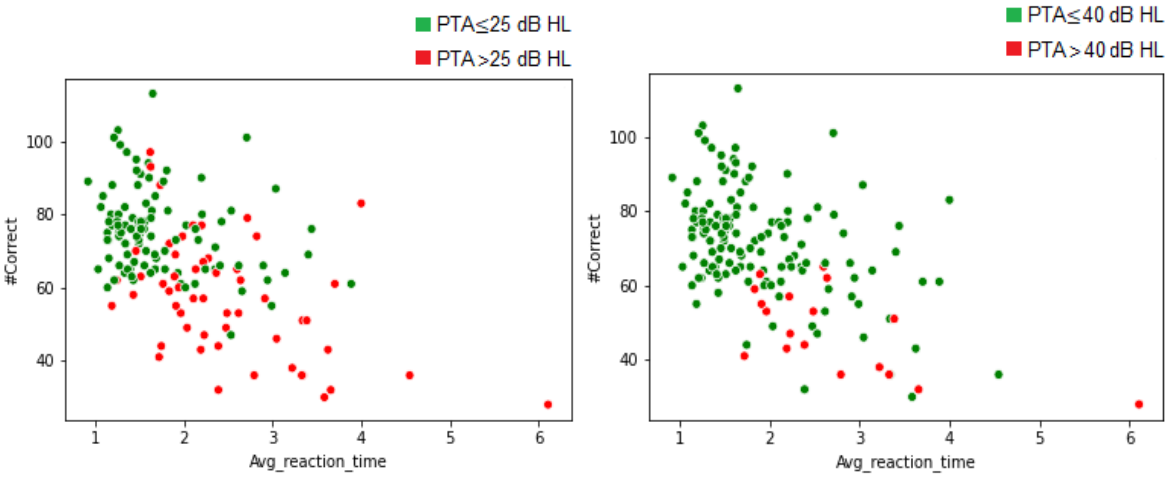


Figure 18. Scatterplots of the average reaction time combined with the number of correct responses for criterion 1 (left panel) and criterion 2 (right panel).

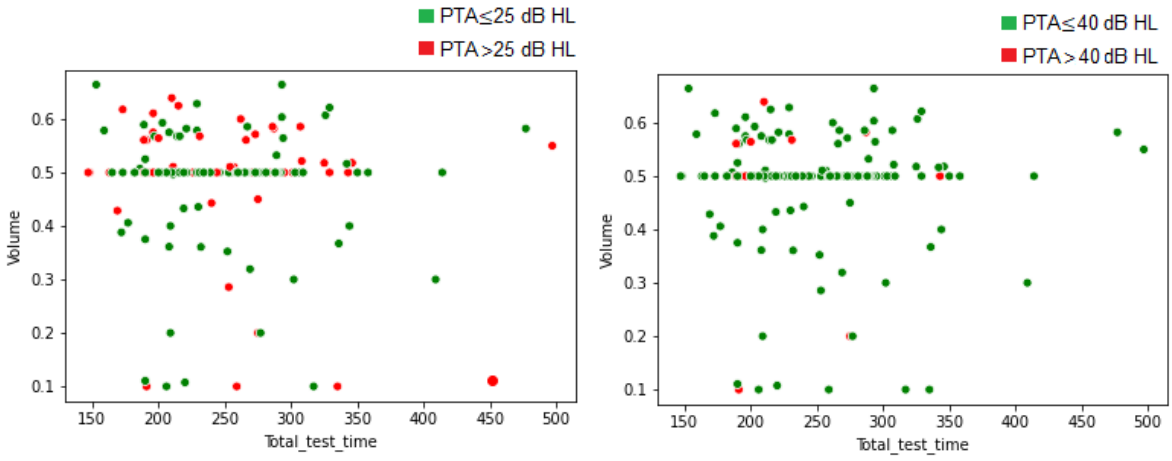


Figure 19. Scatterplots of the total test duration combined with the volume for criterion 1 (left panel) and criterion 2 (right panel).

4.3 Statistical characterization of features as predictors of hearing loss

4.3.1 Evaluation of SRT and age as predictors of PTA outcome

A Generalized Linear Model (GLM) has been used to assess the possible relationship between SRT and PTA.

The binary result of the PTA derived from the WHO criteria for hearing impairment was considered as outcome variable of the model, whereas age, in addition to SRT, were considered as predictors. Also, the interaction between SRT and age was considered.

The logit function was used as link function of the model, since the outcome variable representing the result of the PTA is binary ('pass'/ 'fail').

Focusing first on a single-predictor GLM, the SRT showed a significant correlation ($p << 0.01$) for each of the two WHO criteria. Considering instead a multiple predictor GLM, the results obtained for the two WHO criteria were different.

SRT and age were both significantly correlated with PTA binary output for criterion 1 ($p < 0.01$) while, concerning criterion 2, only SRT was a significant predictor (SRT: $p << 0.01$, Age: $p = 0.1$).

Lastly, the GLM including the interaction between SRT and age showed no significant contribution in predicting the level of hearing loss ($p = 0.4$ for criterion 1 and $p = 0.07$ for criterion 2).

4.3.2 ROC curves

To evaluate if SRTs extracted by the proposed speech-in-noise test could predict the outcomes of PTA according to WHO criteria for "slight/mild" (criterion 1) and for "moderate" (criterion 2) hearing impairment, receiver operating characteristic (ROC) curves were built.

The best candidate SRT cut-off for the test was selected from the ROC curve as the coordinates closest to the ideal (0,1) point. Figure 20 shows the ROC curves for each criterion and the associated SRT cut-off.

The area under the ROC curve (AUC), sensitivity, specificity and the test accuracy were computed for both criterion 1 and criterion 2 and reported in Table 5.

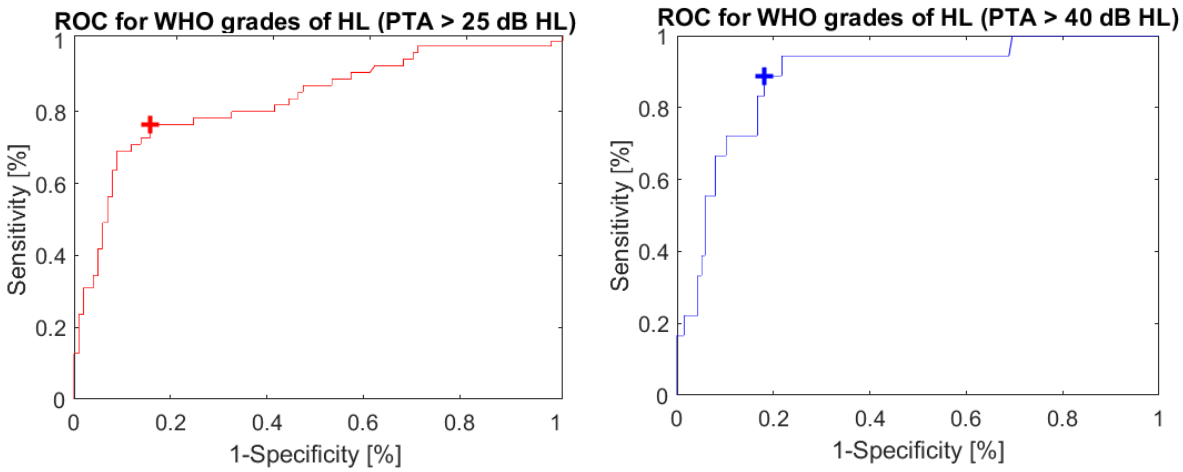


Figure 20. ROC curves. The left panel is related to criterion 1 and the right one to criterion 2. The cross on each graph represents the point associated to the candidate SRT cut-off.

	Criterion 1	**Criterion 2**
$SRT_{cut\text{-}off}$ [dB SNR]	-8.875	-6
AUC	0.83	0.89
Sensitivity [%]	76	89
Specificity [%]	84	81
Accuracy [%]	81	83

Table 5. Values of SRT cut-off, AUC, sensitivity, specificity, and accuracy for the two WHO criteria.

4.3.3 Evaluation of other test variables as predictor of PTA outcome

To further analyze the capability of the proposed speech-in-noise test to predict the PTA results derived from the WHO criteria for mild and moderate hearing impairment, also other variables extrapolated during the test have been considered.

Thus, three different GLMs have been computed for each feature, using first the binary PTA results referred to criterion 1 as response variable y, then the one related to criterion 2.

Since the dependent variable of the model has a binomial distribution, the link function used is ‘logit’.

Hence, the GLMs implemented were:

– single predictor GLM: logit(y) ~ 1 + x1, with x1 being the feature considered
– GLM considering x1 and SRT as predictors: logit(y) ~ 1 + x1 + x2
– GLM also considering the interaction between x1 and SRT:
  logit(y) ~ 1 + x1 + x2+x1*x2

The analysis outcomes are reported respectively in Table 6 (criterion 1) and Table 7 (criterion 2). The adjusted R-squared is reported for each model analyzed, showing the amount of variance in the response variable that is explained by the considered independent variables.

	Single predictor GLM		**Multiple predictor GLM**		**GLM with interactions**	
	R^2_{adj}	p-value	R^2_{adj}	p-value	R^2_{adj}	p-value
SRT	0.34	1.99e-08				
Age	0.40	2.03e-08	0.44	1.33e-05, 0.01	0.44	0.002, 0.74, 0.41
Score	0.16	1.27e-05	0.39	0.001,3.56e-07	0.40	0.16, 5.33e-06, 0.15
# Trials	0.32	5.16e-08	0.37	0.01, 0.001	0.37	0.06, 0.41, 0.94
# Correct Answers	0.35	2.13e-08	0.38	0.006, 0.005	0.38	0.04, 0.44, 0.96
%Correct Answers	0.3	1.06e-06	0.35	0.12, 0.08	0.34	0.16, 0.5, 0.4
Total_test_time [s]	0	0.95	0.33	0.61, 1.74e-08	0.37	0.07, 0.0003,0.01
Avg_reaction_time [s]	0.19	6.19e-07	0.35	0.02, 7.45e-06	0.35	0.36, 0.0005, 0.098
Volume	0	0.47	0.33	0.56, 2.12e-08	0.33	0.79, 0.13, 0.52

Table 6. R squared adjusted and p-value of each predictor for 3 different kind of GLM, selecting as response class the binary vector ('pass'/'fail') related to criterion 1.

	Single predictor GLM		Multiple predictor GLM		GLM with interactions	
	R^2_{adj}	p-value	R^2_{adj}	p-value	R^2_{adj}	p-value
SRT	0.23	7.62e-07				
Age	0.13	4.5e-04	0.24	0.1, 6.1e-04	0.28	0.37, 0.03, 0.07
Score	0.14	3.03e-05	0.33	0.002, 9.57e-06	0.32	0.003, 6.82e-05, 0.09
# Trials	0.24	1.36e-06	0.25	0.03, 0.10	0.33	0.02, 0.02, 0.01
# Correct Answers	0.26	7.24e-07	0.26	0.023, 0.45	0.30	0.04, 0.03, 0.01
%Correct Answers	0.15	0.06	0.22	0.37, 6.9e-04	0.22	0.73, 0.97, 0.80
Total_test _time [s]	0.01	0.13	0.23	0.25, 8.13e-07	0.24	0.31, 0.81, 0.43
Avg_reaction_time [s]	0.07	7e-04	0.23	0.82,1e-04	0.22	0.91,0.004, 0.27
Volume	0	0.97	0.22	0.74, 8.39e-07	0.28	0.91,0.01,0.02

Table 7. R squared adjusted and p-value of each predictor for 3 different kind of GLM, selecting as response class the binary vector ('pass'/'fail') related to criterion 2.

In general, adding new predictors to the model, bring to a higher value of R^2_{adj}.

Focusing on the single predictor model for both criteria, all the features except the total test duration and the volume showed a statistically significant association ($p<0.05$) with the response variable, therefore, changes in these features contribute in a certain amount to the variation of the response. Instead, the screening result could be considered independent on the set volume level. The distribution of this variable was in fact very narrow, centered around the default value and for this reason variations in volume were minimal and did not affect the test result.

Considering the multiple-predictor GLM for criterion 1, a significant contribution has been found, for both predictors, for age, score, number of trials, number of correct responses and average reaction time, whereas in all the other attempts, concerning the percentage of correct responses, the total test duration, and the volume, only SRT showed a p-value lower than the significance level.

Analyzing instead the same type of model, but this time related to criterion 2, only the model considering score and SRT showed a statistical significance for both predictors.

The models computed using respectively the age and the SRT, the percentage of correct responses and the SRT, the total test duration and the SRT and the volume and the SRT showed only a statistically significant association for what regards SRT. All the other models, considering the remaining features, had the opposite behavior, with the p-value related to SRT above the significance level and the other predictor showing a meaningful association with the response variable.

As regards the models with the interactions, instead, the 2 criteria showed different results.

Concerning criterion 1, no model showed significant relationships for all three components (x1, SRT and interaction), while models involving '#Trials', '%Correct' and 'Volume' had no significant predictors. Only the interaction between the total test duration and SRT was significant.

A longer duration of the test may be due to a better hearing, when the subject performs well in the test (i.e., lower SRT) and therefore listens to more VCV proposals before finishing the test. However, even when the subject has poorer abilities to recognize speech in noise (i.e., higher SRT) the time spent on the test may be high; although the subjects goes through a lower number of trials, they will struggles to distinguish stimuli and therefore they will respond with a higher indecision to each single proposal of the stimulus.

Therefore, in presence of a significant interaction term, the effect of one predictor variable on the dependent variable changes at different values of the other predictor variable, indeed, the effect of the test duration on the result of the screening is different for different values of SRT.

Considering instead criterion 2, both the model involving the number of trials and the one involving the number of correct responses resulted to be significant for all the three predictors. The model with the percentage of correct responses and the total test duration had instead no meaningful components, while the interactions were statistically relevant only for the number of trials, the number of correct responses and the volume.

4.4 Classification

The main aim of the last part of the study was to classify observations into ‘pass’ or ‘fail’ using a machine learning approach based on the gathered data in order to predict the result of the PTA in terms of level of hearing impairment. The main interest of a screening test is to make the subject aware of the problem before it leads to serious consequences. Therefore, they are mainly addressed to identify ‘slight/mild’ hearing loss.

This reason, together with the fact that the number of subjects with ‘moderate’ hearing impairment in the sample was very low (only 18 observations), led to the decision to investigate the classification performances only for criterion 1.

Each classifier has been evaluated by feeding it with different combinations of the available features:

- the full set of features.
- a subset of nine features: ‘SRT’, ‘Score’, ‘Age’, ‘#Trials’, ‘#Correct’, ‘%Correct’, ‘Avg_reaction_time’, ‘Total_test_time’, ‘Volume’.
- a subset of nine features: ‘SRT’, ‘Age’, ‘#Trials’, ‘#Correct’, ‘%Correct’, ‘Avg_reaction_time’, ‘Total_test_time’, ‘Volume’, ‘Score_classes’.
- a subset of eight features: ‘SRT’, ‘Age’, ‘#Trials’, ‘#Correct’, ‘%Correct’, ‘Avg_reaction_time’, ‘Total_test_time’, ‘Volume’.
- a subset of seven features: ‘SRT’, ‘Score’, ‘Age’, ‘#Trials’, ‘#Correct’, ‘%Correct’, ‘Avg_reaction_time’.
- a subset of six features: ‘SRT’, ‘Age’, ‘#Trials’, ‘#Correct’, ‘%Correct’, ‘Avg_reaction_time’.
- a subset of four features: ‘SRT’, ’Age’, ‘#Correct’ and ‘Avg_reaction_time’.

The different subgroups of variables have been chosen accordingly to the results found in Section 4.2 and 4.3.3. Indeed, the volume and the total test duration were discarded as they present no significant correlation with respect to PTA and therefore a change in these variables should not affect the screening result. Instead, the two variables related to the score of the questionnaire were rejected because of some critical aspects (i.e., language dependency and subjective component) that will be further analyzed in the ‘Discussion’ chapter. Lastly, one feature among the number of trials, the number of correct responses and

the percentage of correct responses, that substantially bring the same information, was chosen.

As previously introduced, the dataset was randomly split into a training and a test partition, with a ratio of 80% (124 ears) and 20% (32 ears) respectively.

The dataset has undergone a scaling procedure to assure that the model features had null mean and unitary variance, then, a 5-fold cross-validation has been introduced.

After fitting, each model was tested on the test set.

The performance for each classification model under study was assessed by measuring the following parameters: accuracy on the training set and on the test set, area under the curve (AUC), specificity, sensitivity, True Positives (TP), False negatives (FN), FNR (False Negative Rate), Precision and F-measure.

Average value and standard deviation of models' performance indicators have been calculated over 1000 iterations to reduce variability due to the reduced size of training and test sets.

The algorithms investigated were Decision Trees (DT), Support Vector Machines (SVM), logistic regression, K-Nearest-Neighbors, ensemble logistic regression, Random forests and Gradient boosting.

4.4.1 Decision Trees

A Decision Tree classification algorithm was considered in the first place as interpretable splitting rules can be extracted from the paths of the tree. The Gini index was chosen as criterion for measuring the impurity of a node.

Figure 21 visualizes the optimal DT obtained using the full set of features available. Each box contains the splitting rule of the node, the relative Gini index, the number of samples at the given node and the related number of samples for each class.

The first splitting rule in the root node takes into account the SRT value, with a cut-off value of -8.19 dB SNR, similar to the one estimated from the ROC curve and described in Section 4.3.2. The SRT value is used again as a splitting rule on several nodes throughout the tree. First on level three, to classify a subset of 46 observations from subjects with an average reaction time lower than 1.82. Then, SRT is present again in the fourth level, in the right branch of the three, creating a partition of 25 samples having made more than 63.49 correct responses and having age higher than 52 y.o. Besides, in the fifth level of depth, SRT creates

a subsample of 7 observations with average reaction time higher than 1.82 s, having tuned the volume to a value higher than 0.29 and having made a number of test trials lower than 72.99. In addition to SRT, also other features contribute to the classification in a consistent way and are present in more than one occasion in the decision making process of the tree; for example the number of correct responses (with 'fail' outcomes associated with lower number of correct responses), the average reaction time and age (with 'fail' associated with older age).

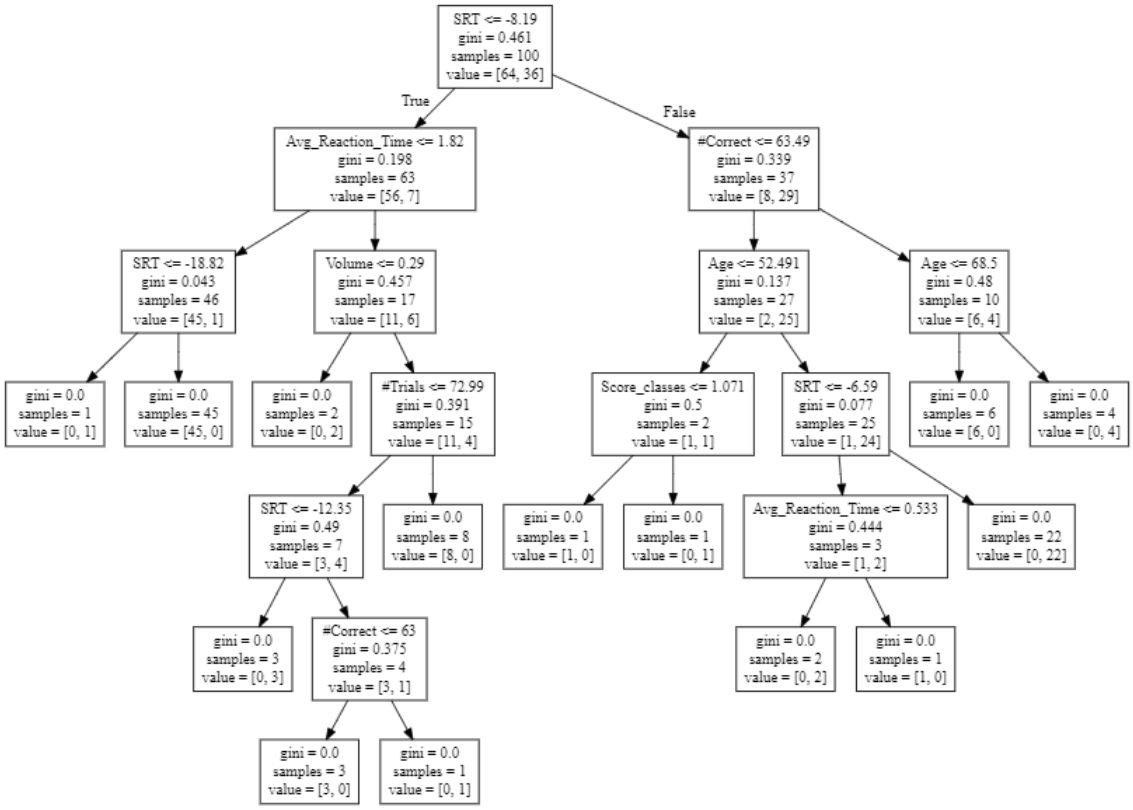


Figure 21. Optimal DT model for classification of ears into 'pass' and 'fail' using the full set of features as input variables and the WHO definition of normal hearing / mild hearing loss as output variable. No constraints on the maximum depth have been defined.

Table 8 displays the performance indicators extracted for each model considering different subgroups of features each time. Test and training accuracy show similar values, around 0.76, AUC is around 0.74, Specificity is quite high, above 0.80, while F-measure is quite low, around 0.66.

In general, there are no differences in the values of the indicators obtained from models built starting from different number of features.

DECISION TREES	ALL	9 FEAT (NO SCORE)	9 FEAT (NO SCORE CLASSES)	8 FEAT	7 FEAT	6 FEAT	4 FEAT
TEST ACCURACY	0.76±0.07	0.77±0.08	0.76±0.07	0.77±0.07	0.76±0.07	0.77±0.07	0.77±0.07
TRAINING ACCURACY	0.76±0.04	0.76±0.04	0.76±0.04	0.76±0.04	0.76±0.03	0.76±0.04	0.78±0.03
AUC	0.74±0.08	0.75±0.08	0.74±0.08	0.74±0.08	0.74±0.07	0.74±0.08	0.75±0.08
Specificity	0.81±0.09	0.82±0.09	0.81±0.09	0.82±0.09	0.81±0.09	0.82±0.09	0.82±0.09
Sensitivity	0.66±0.15	0.68±0.15	0.67±0.15	0.67±0.15	0.67±0.14	0.67±0.14	0.69±0.14
TP	7.31±1.63	7.45±1.63	7.32±1.66	7.35±1.6	7.38±1.57	7.38±1.59	7.59±1.52
FN	3.69±1.63	3.55±1.63	3.68±1.66	3.65±1.6	3.62±1.57	3.62±1.59	3.42±1.52
FNR	0.34±0.15	0.32±0.15	0.33±0.15	0.33±0.15	0.33±0.14	0.33±0.14	0.31±0.14
Precision	0.66±0.12	0.67±0.13	0.66±0.12	0.67±0.12	0.66±0.11	0.67±0.12	0.68±0.12
F-measure	0.65±0.11	0.66±0.11	0.65±0.11	0.66±0.11	0.65±0.1	0.66±0.1	0.68±0.1

Table 8. Classification performance and variability of performance of the DT models with different input features.

Let's now focus on the model based on the smaller set of features; the corresponding optimal DT model is represented in Figure 22.

The first splitting rule in the root node, takes into account the SRT value, with a threshold value of -8.873 dB SNR, very similar to the cut-off value estimated from the ROC curve (-8.875 dB SNR) for criterion 1. The SRT value is used again as a splitting rule in multiple occasions. First on level four, to classify a subset of 6 observations from subjects younger than 68 years that have obtained a number of correct responses in the test lower than 61.49. Moreover, SRT is present again in the fifth level, in the left branch of the three, creating a partition of 20 samples having made more than 58.99 correct responses, having an average reaction time higher than 1.62 and age higher than 45 y.o. Furthermore, starting from this sample of 20 observations, SRT creates a subsample of 4 (with average reaction time < 2.2 s) and finally, in the last level of rules, creates a group of three observations with a number of correct responses lower than 84.5 and an average reaction time of 1.86 s.

In addition to it, also other features are used to classify observations into pass and fail and are present multiple times going down the tree. In particular, 'fail' observations are mostly

associated to high values of average reaction time and age and a lower number of correct responses.

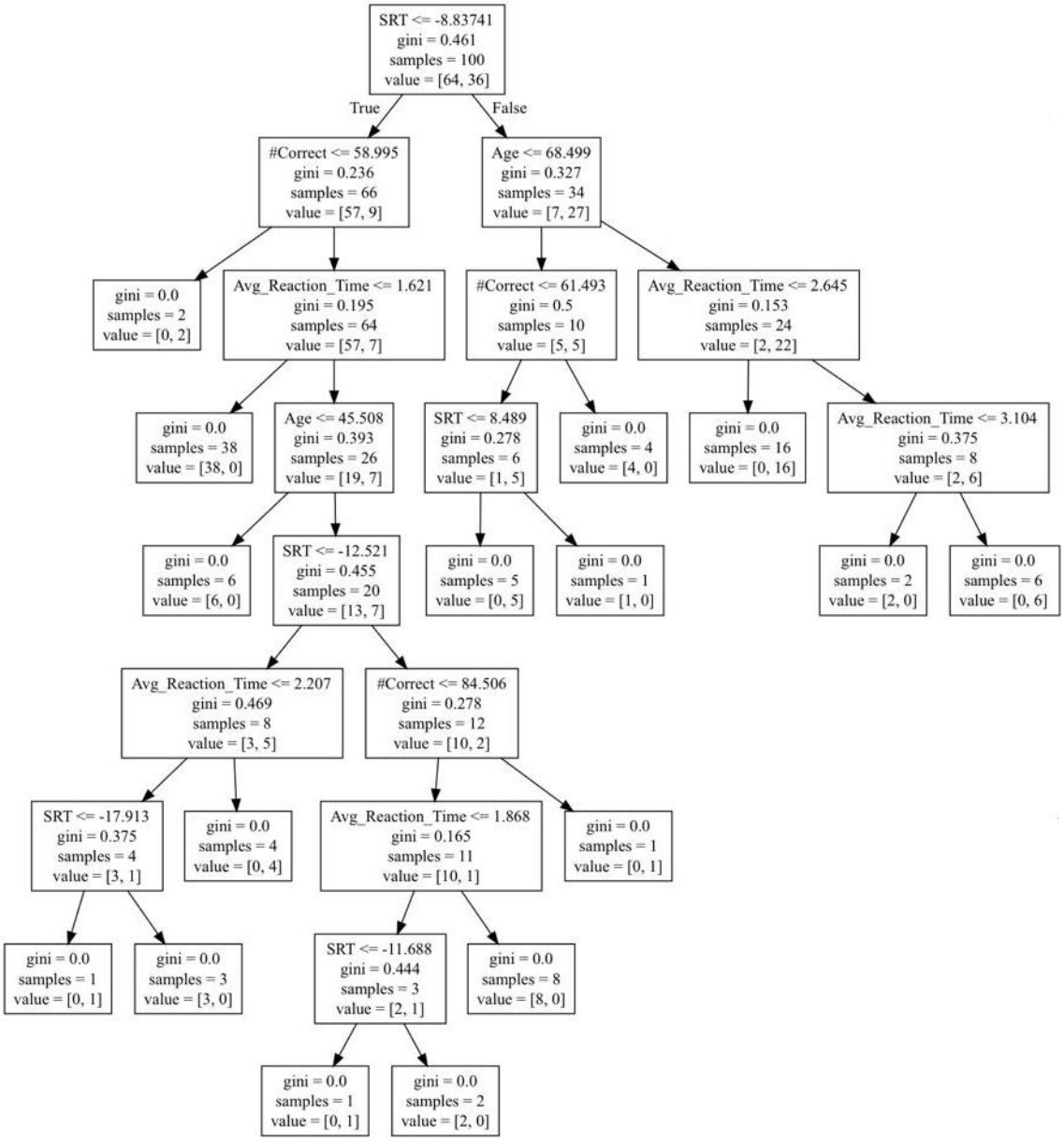


Figure 22. Optimal DT model for classification of ears into 'pass' and 'fail' using only four features as input variables ('SRT', 'Age', '#Correct' and 'Avg_reaction_time') and the WHO definition of normal hearing / mild hearing loss as output variable. No constraints on the maximum depth have been defined.

In order to prevent overfitting due to a reduced number of features and obtain less complex decision rules, different depths (3, 4 and 5) were investigated for the trees. Results are reported in Table 9.

	depth=3	depth=4	depth=5
TEST ACCURACY	0.77±0.07	0.77±0.07	0.78±0.07
TRAINING ACCURACY	0.78±0.04	0.78±0.04	0.78±0.04
AUC	0.82±0.07	0.8±0.08	0.79±0.08
SPECIFICITY	0.78±0.1	0.8±0.09	0.81±0.09
SENSITIVITY	0.76±0.14	0.73±0.14	0.72±0.15
TP	8.36±1.56	8.02±1.58	7.87±1.62
FN	2.65±1.56	2.98±1.58	3.13±1.62
FNR	0.24±0.14	0.27±0.14	0.28±0.15
PRECISION	0.66±0.11	0.67±0.11	0.67±0.12
F-MEASURE	0.7±0.09	0.69±0.1	0.68±0.11

Table 9. Classification performance and variability of performance of the DT models with three different limitations on the maximum depth achievable by the tree.

Adding a limit to depth has led to an improvement in AUC ($\simeq 0.86$), sensitivity ($\simeq 0.74$), TP, FN and FNR. The other parameters remain almost unchanged.

The DT with four selected features and three levels of depth is displayed in Figure 23.

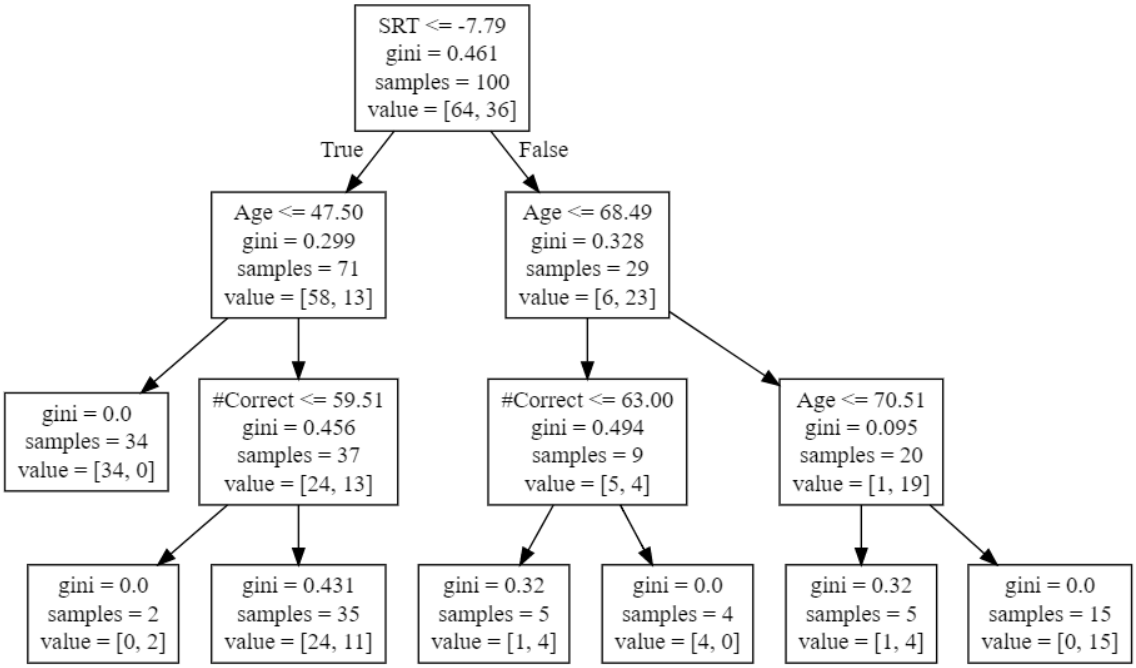


Figure 23. Optimal DT model for classification of ears into 'pass' and 'fail' using only four features as input variables ('SRT', 'Age', '#correct' and 'Avg_reaction_time') and the WHO definition of normal hearing / mild hearing loss as output variable. Maximum depth has been set equal to 3.

As for the previous DTs, the decision rule of the top node is based on the SRT, presenting a cut-off equal to -7.79 dB SNR. Also age has an important contribution in the decision rules, indeed is present in both branches in the second level and in correspondence of the right branch in the third level. The number of correct responses is present as the discrimination variable in the third level, in both branches. Even in this case 'fail' records are associated with older age and a lower number of correct responses. Considering only three levels of depth, none of the splitting rules involves the average reaction time.

4.4.2 Support vector machines

Table 10 shows the classification performance and the variability of performance obtained with SVM, considering different subsets of features.

It can be noticed that most of the parameters measured presented better value with respect to the previous machine learning approach; for example, the test accuracy is around 0.80, the training accuracy exceeds 0.80, AUC is around 0.90, F-measure is higher than 0.7. Specificity and precision show comparable values between the two approaches.

The SVM method presents similar accuracy values, with training accuracy slightly higher with respect to the test accuracy.

As for the DTs, there are no significant differences between the result obtained starting with a different subset of features.

SVM	ALL	9 FEAT (NO SCORE)	9 FEAT (NO SCORE CLASSES)	8 FEAT	7 FEAT	6 FEAT	4 FEAT
TEST ACCURACY	0.82±0.07	0.8±0.06	0.81±0.06	0.78±0.07	0.81±0.06	0.78±0.07	0.78±0.07
TRAINING ACCURACY	0.83±0.02	0.82±0.02	0.83±0.02	0.8±0.02	0.83±0.02	0.81±0.02	0.81±0.02
AUC	0.91±0.05	0.89±0.05	0.91±0.05	0.88±0.06	0.91±0.05	0.88±0.06	0.88±0.06
Specificity	0.82±0.09	0.8±0.09	0.81±0.09	0.79±0.09	0.81±0.09	0.78±0.1	0.78±0.1
Sensitivity	0.81±0.12	0.8±0.13	0.81±0.13	0.78±0.13	0.81±0.14	0.78±0.14	0.78±0.14
TP	8.95±1.32	8.81±1.41	8.89±1.41	8.53±1.42	8.95±1.5	8.64±1.51	8.64±1.51
FN	2.05±1.32	2.19±1.41	2.11±1.41	2.47±1.42	2.05±1.5	2.37±1.51	2.37±1.51
FNR	0.19±0.12	0.2±0.13	0.19±0.13	0.22±0.13	0.19±0.14	0.21±0.14	0.21±0.14
Precision	0.72±0.11	0.69±0.1	0.7±0.1	0.67±0.1	0.7±0.11	0.67±0.1	0.67±0.1
F-measure	0.75±0.08	0.74±0.08	0.74±0.09	0.71±0.09	0.74±0.08	0.71±0.09	0.71±0.09

Table 10. Classification performance and variability of performance of the SVM models with different input features.

4.4.3 Logistic regression

The third methodology investigated was logistic regression. The corresponding results are described in Table 11. The performances obtained with this approach are fully comparable with those obtained with SVM, with the test accuracy, specificity, and sensitivity around 0.80, the training accuracy exceeding 0.80, AUC around 0.90, and F-measure is higher than 0.72. Also in this case, changing the number of features hasn't changed consistently the performance of the models.

LOGISTIC REGRESSION	ALL	9 FEAT (NO SCORE)	9 FEAT (NO SCORE CLASSES)	8 FEAT	7 FEAT	6 FEAT	4 FEAT
TEST ACCURACY	0.82±0.07	0.8±0.06	0.81±0.06	0.79±0.06	0.81±0.06	0.79±0.07	0.79±0.07
TRAINING ACCURACY	0.83±0.02	0.81±0.02	0.83±0.02	0.8±0.02	0.83±0.02	0.8±0.02	0.8±0.02
AUC	0.91±0.05	0.9±0.05	0.91±0.05	0.89±0.05	0.91±0.05	0.9±0.05	0.9±0.05
Specificity	0.82±0.09	0.8±0.09	0.82±0.09	0.8±0.09	0.82±0.09	0.79±0.09	0.79±0.09
Sensitivity	0.8±0.13	0.79±0.13	0.8±0.13	0.79±0.13	0.81±0.13	0.8±0.13	0.8±0.13
TP	8.8±1.38	8.72±1.41	8.76±1.39	8.67±1.38	8.87±1.45	8.83±1.45	8.83±1.45
FN	2.2±1.38	2.28±1.41	2.25±1.39	2.34±1.38	2.13±1.45	2.18±1.45	2.18±1.45
FNR	0.2±0.13	0.21±0.13	0.2±0.13	0.21±0.13	0.19±0.13	0.2±0.13	0.2±0.13
Precision	0.72±0.11	0.69±0.1	0.71±0.1	0.68±0.1	0.71±0.1	0.68±0.1	0.68±0.1
F-measure	0.75±0.09	0.73±0.08	0.74±0.08	0.72±0.08	0.75±0.08	0.73±0.09	0.73±0.09

Table 11. Classification performance and variability of performance of the logistic regression models with different input features.

4.4.4 K-Nearest Neighbor

In a KNN classifier, each new observation, represented as a point in the features space, is assigned to the class related to the majority of its k nearest neighbors. Therefore, the two main parameters to set were the number of neighbors k to consider and the metric used to calculate the distance. Throughout this study, a number of neighbors k=5 (found using the elbow method) and Euclidean distance metric have been chosen.

Outcomes related to the models implemented are reported in Table 12.

The performances are not as good as for SVM and logistic regression, except for TP, FN and FNR. Though, the results found are better with respect to DTs in terms of AUC (≃ 86%), sensitivity (≃0.77) and F-measure (≃0.7). Training accuracy and precision are similar to what has been found for DT, while specificity is a bit lower (≃0.77).

Accuracy values are comparable, with training accuracy slightly higher than the test accuracy.

As in the previous cases, building models from different subsets of features does not have any particular effect on the classification performance.

KNN	ALL	9 FEAT (NO SCORE)	9 FEAT (NO SCORE CLASSES)	8 FEAT	7 FEAT	6 FEAT	4 FEAT
TEST ACCURACY	0.77±0.07	0.76±0.07	0.78±0.07	0.75±0.07	0.79±0.06	0.77±0.07	0.77±0.07
TRAINING ACCURACY	0.78±0.03	0.78±0.03	0.79±0.03	0.77±0.03	0.81±0.03	0.79±0.03	0.79±0.03
AUC	0.85±0.06	0.85±0.06	0.86±0.06	0.85±0.06	0.87±0.06	0.87±0.06	0.87±0.06
Specificity	0.77±0.1	0.77±0.1	0.77±0.1	0.74±0.1	0.79±0.1	0.77±0.11	0.77±0.11
Sensitivity	0.77±0.14	0.76±0.15	0.78±0.13	0.78±0.14	0.78±0.14	0.76±0.15	0.76±0.15
TP	8.5±1.56	8.4±1.6	8.61±1.44	8.55±1.54	8.62±1.51	8.41±1.62	8.41±1.62
FN	2.5±1.56	2.6±1.6	2.39±1.44	2.45±1.54	2.38±1.51	2.59±1.62	2.59±1.62
FNR	0.23±0.14	0.24±0.15	0.22±0.13	0.22±0.14	0.22±0.14	0.24±0.15	0.24±0.15
Precision	0.65±0.1	0.65±0.1	0.66±0.1	0.62±0.1	0.68±0.1	0.66±0.11	0.66±0.11
F-measure	0.7±0.09	0.69±0.09	0.71±0.09	0.68±0.09	0.72±0.08	0.69±0.09	0.69±0.09

Table 12. Classification performance and variability of performance of the KNN models with different input features.

4.4.5 Ensemble logistic regression

The first ensemble technique investigated was a ensemble logistic regression method, whose aim is to combine the four machine learning algorithms previously considered into a unique model in order to try to outstand each of these algorithm considered alone in terms of performances. Table 13 shows the obtained results.

The training accuracy is close to 1 (0.95) and presents far higher values with respect to the test accuracy (0.78), clear symptom of overfitting.

The measured test accuracy is similar to the one found for DT and logistic regression. AUC and specificity are quite high ($\simeq$0.89 and $\simeq$0.82 respectively), whereas sensitivity is lower ($\simeq$0.69).

Results are similar even considering models based on different features.

ENSEMBLE LOG-REG	ALL	9 FEAT (NO SCORE)	9 FEAT (NO SCORE CLASSES)	8 FEAT	7 FEAT	6 FEAT	4 FEAT
TEST ACCURACY	0.78±0.07	0.78±0.07	0.78±0.07	0.78±0.07	0.77±0.07	0.78±0.07	0.78±0.07
TRAINING ACCURACY	0.95±0.02	0.95±0.02	0.95±0.02	0.95±0.02	0.95±0.02	0.95±0.02	0.95±0.02
AUC	0.89±0.05	0.88±0.06	0.89±0.05	0.88±0.06	0.89±0.05	0.89±0.05	0.89±0.05
Specificity	0.82±0.09	0.82±0.09	0.82±0.09	0.83±0.09	0.81±0.09	0.83±0.09	0.83±0.09
Sensitivity	0.69±0.15	0.69±0.15	0.69±0.15	0.68±0.14	0.69±0.14	0.7±0.14	0.7±0.14
TP	7.58±1.61	7.58±1.6	7.55±1.62	7.47±1.57	7.62±1.55	7.71±1.51	7.71±1.51
FN	3.42±1.61	3.43±1.6	3.45±1.62	3.53±1.57	3.38±1.55	3.29±1.51	3.29±1.51
FNR	0.31±0.15	0.31±0.15	0.31±0.15	0.32±0.14	0.31±0.14	0.3±0.14	0.3±0.14
Precision	0.68±0.12	0.68±0.12	0.68±0.12	0.68±0.12	0.67±0.11	0.7±0.12	0.7±0.12
F-measure	0.68±0.11	0.68±0.11	0.67±0.11	0.67±0.1	0.67±0.1	0.69±0.1	0.69±0.1

Table 13. Classification performance and variability of performance of the ensemble logistic regression models with different input features.

4.4.6 Random Forests

Another family of classification methods considered is random forest, where the output of the model is obtained by averaging or voting, starting from the combination of a multitude of single decision trees randomly created. The number of trees in the forest was varied between 10, 50 and 100 trees. Results are shown in Table 14. The Random forest approach improves with respect to DTs in terms of all the parameters considered especially AUC (0.8), sensitivity (≃0.73), TP, precision (>0.7), and F-measure. There are no big differences in terms of accuracy.

Changing the number of features considered and the number of trees in the forest does not lead to any significant change in the performance other than the extension of the execution times due to a higher number of trees.

Indeed, the computational time required for the training of a forest with only 10 trees is almost the double of the time required for a single decision tree, which is quite fast. On the

other hand, increasing the number of trees bring to a consistent increase of the time cost, with respect to the use of a single tree.

RANDOM FOREST	**ALL FEATURES**			**4 FEATURES**		
N° OF TREES	10	50	100	10	50	100
TEST ACCURACY	0.79±0.06	0.8±0.06	0.8±0.06	0.8±0.07	0.8±0.06	0.81±0.06
TRAINING ACCURACY	0.8±0.03	0.8±0.02	0.8±0.02	0.8±0.03	0.8±0.03	0.81±0.03
AUC	0.87±0.06	0.89±0.05	0.89±0.05	0.88±0.06	0.88±0.06	0.9±0.05
Specificity	0.82±0.09	0.84±0.09	0.85±0.09	0.83±0.09	0.84±0.09	0.86±0.09
Sensitivity	0.73±0.14	0.71±0.16	0.71±0.16	0.74±0.14	0.72±0.14	0.71±0.15
TP	8.02±1.58	7.81±1.73	7.85±1.74	8.12±1.54	8.01±1.55	7.85±1.65
FN	2.99±1.58	3.19±1.73	3.15±1.74	2.88±1.54	2.99±1.55	3.15±1.65
FNR	0.27±0.14	0.29±0.16	0.29±0.16	0.26±0.14	0.27±0.14	0.29±0.15
Precision	0.7±0.11	0.73±0.12	0.73±0.12	0.71±0.12	0.73±0.14	0.75±0.12
F-measure	0.7±0.09	0.7±0.1	0.7±0.1	0.71±0.09	0.7±0.1	0.72±0.1
TOTAL RUNNING TIME	≃ 2*k	≃ 10*k	≃ 20*k	≃ 2*k	≃ 10*k	≃ 20*k

Table 14. Classification performance and variability of performance of the optimal DT models with all (left) and with four (right) input features. Results for forests with 10, 50 and 100 are reported. The execution time is approximated as multiple of the time required for the training of a decision tree (k).

When more decision trees are brought together to form a forest, it becomes very hard to understand the mechanisms underneath classification. Despite so, the relative contribution of each single feature of the model to the process of classification can be analyzed.

Figure 24 displays the relative importance scores for each input feature, respectively for the forest considering all the features and the one considering only 4 features.

It can be noticed that SRT and age both play an important role in the decision making process of the forest, whereas the total test duration, the volume and the score of the questionnaire represented by three classes, which indeed are the variable less correlated with the PTA values, do not seem to contribute that much to the classification process.

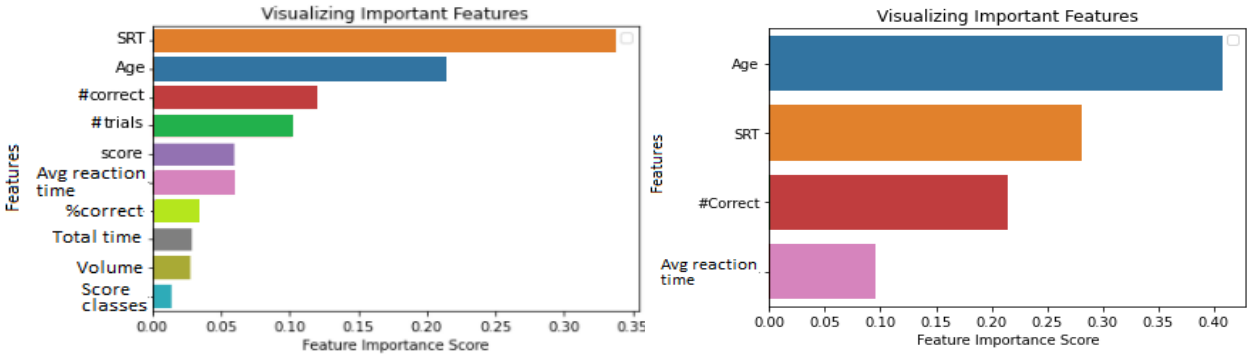


Figure 24. Relative importance scores for each input feature of the model, respectively for the forest considering all the features (left panel) and the one considering only 4 features (right panel).

4.4.7 Gradient Boosting

The last method investigated is gradient boosting, an additive ensemble method which consists in the correction and improvement of a base model (i.e., decision tree) by the successive addition of other decision trees.

In order to tune the model, the Grid Search method has been implemented. The selected parameters for the optimized model were:

- learning rate=0.1, determining the output of each single tree to the final output of the model.
- maximum depth of the individual regression estimators (limiting the number of nodes in the tree) = 3.

– number of 'boosting stages' i.e. the number of trees to be modeled sequentially = 500.
– fraction of the total number of observations to be selected to fit each single tree = 0.7.

Results for both non-optimized and optimized strategy are shown in Table 15 for the models based on all the features and only on 4 features.

GRADIENT BOOSTING	**All features**		**4 features**	
		Optimization		Optimization
TEST ACCURACY	0.78±0.07	0.8±0.06	0.79±0.07	0.79±0.06
TRAINING ACCURACY	0.79±0.03	0.8±0.02	0.79±0.03	0.79±0.03
AUC	0.87±0.06	0.89±0.05	0.88±0.06	0.88±0.05
Specificity	0.82±0.11	0.84±0.09	0.83±0.10	0.82±0.12
Sensitivity	0.7±0.17	0.71±0.16	0.71±0.17	0.72±0.17
TP	7.75±1.87	7.81±1.73	7.78±1.84	7.9±1.97
FN	3.25±1.87	3.19±1.73	3.22±1.84	3.1±1.97
FNR	0.3±0.17	0.29±0.16	0.29±0.17	0.28±0.18
Precision	0.7±0.12	0.73±0.12	0.71±0.13	0.71±0.13
F-measure	0.68±0.11	0.7±0.1	0.69±0.11	0.7±0.1
TOTAL RUNNING TIME	≃ 30*k	≃ 40*k	≃ 30*k	≃ 40*k

Table 15. Classification performance and variability of performance of Gradient Boosting models with all and with only four input features. Results for attempts without and with optimization are reported. The execution time is approximated as multiple of the time required for the training of a decision tree (k).

Even in this case, the performance measured is better than the one of the DT approach, showing comparable results with respect to the previous ensemble method considered, however, the time required to fit the model is actually very long with respect to that of a single tree.

4.4.8 Comparison between classification algorithms

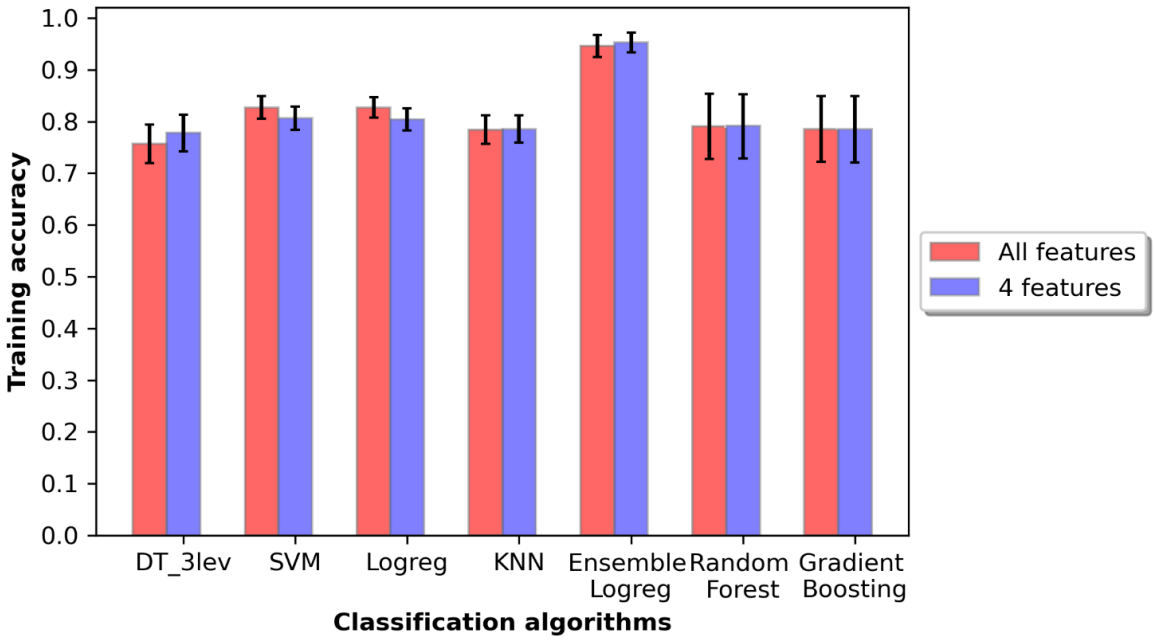


Figure 25. Training accuracies for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

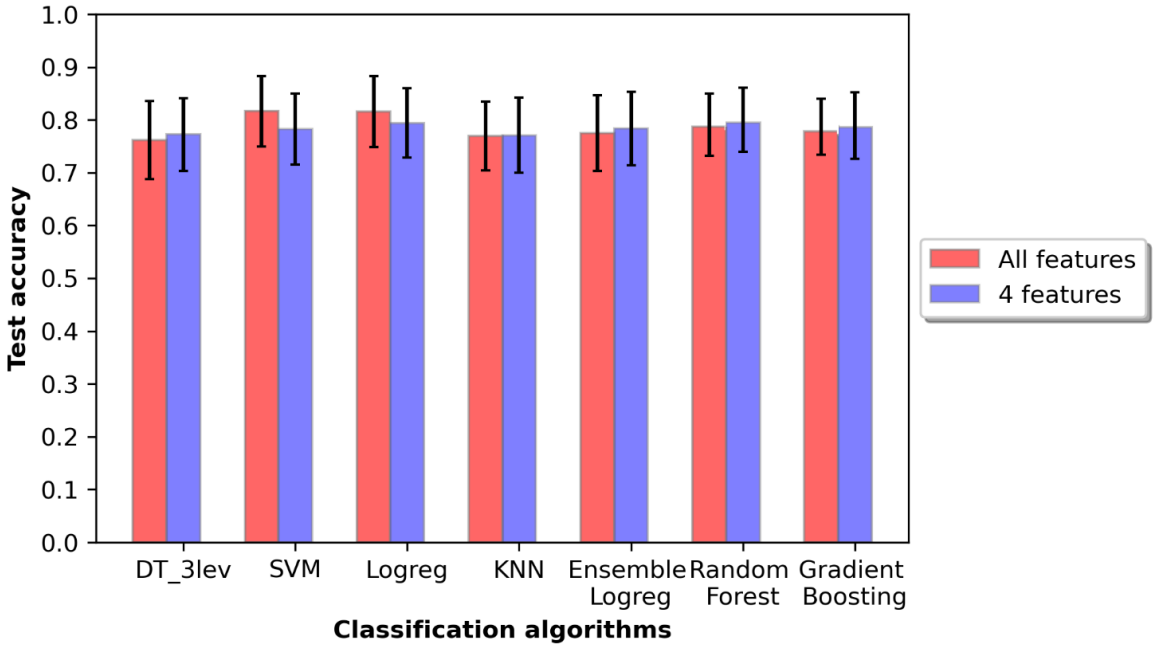


Figure 26. Test accuracies for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

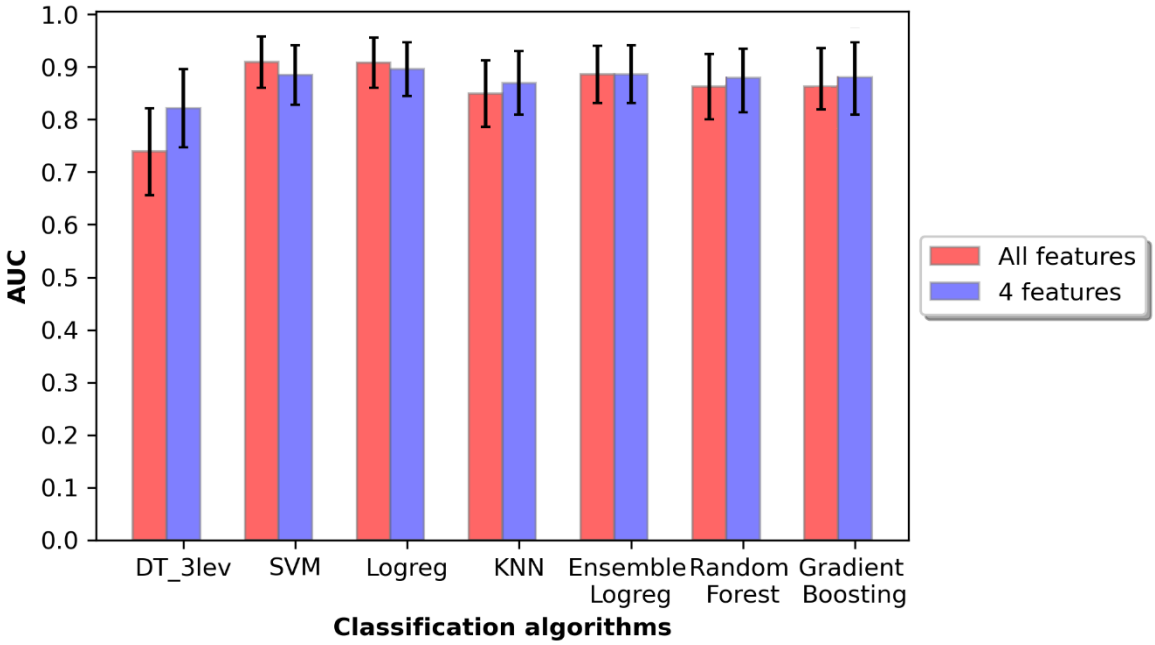


Figure 27. Area under the ROC curve (AUC) for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

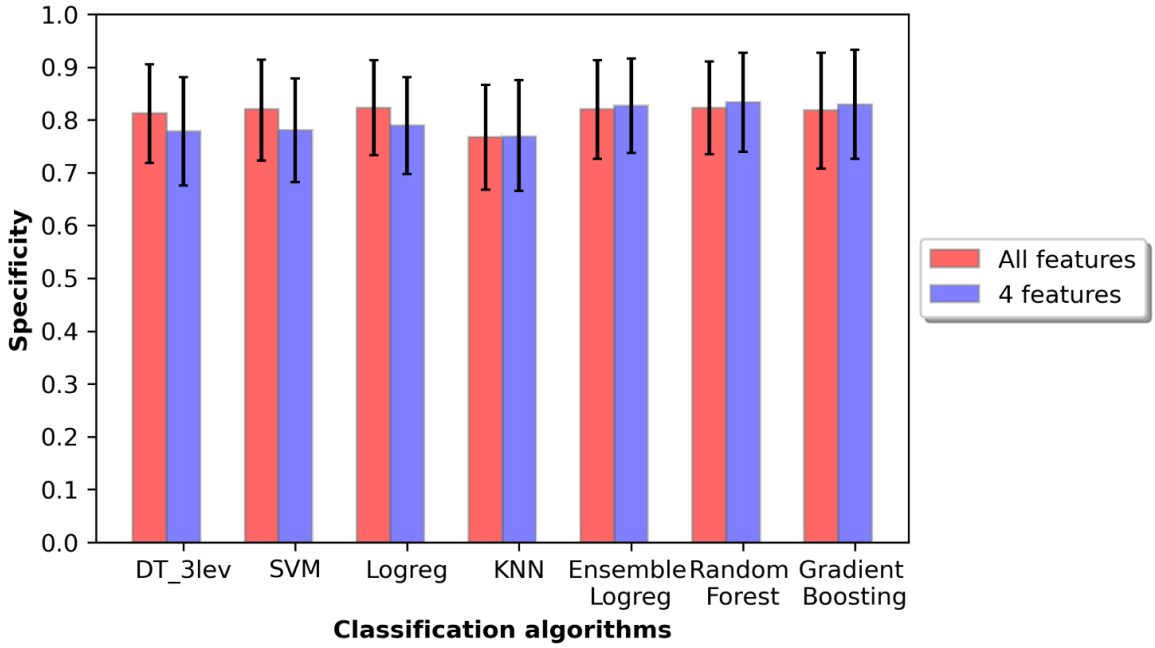


Figure 28. Specificities for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

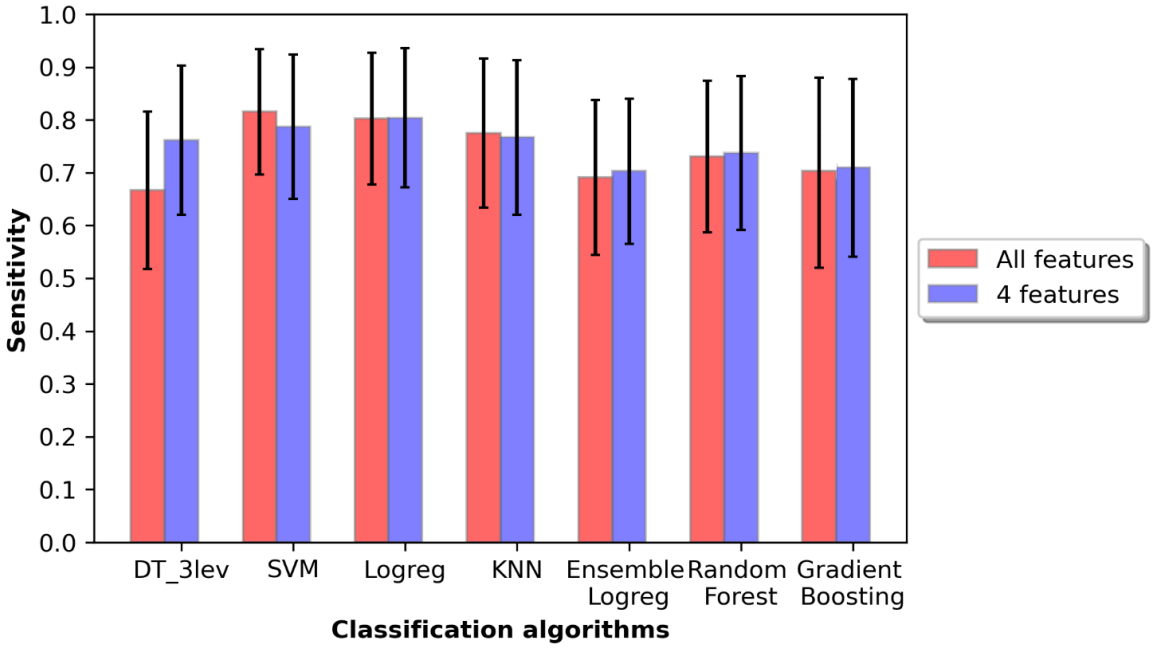


Figure 29.Sensitivities for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

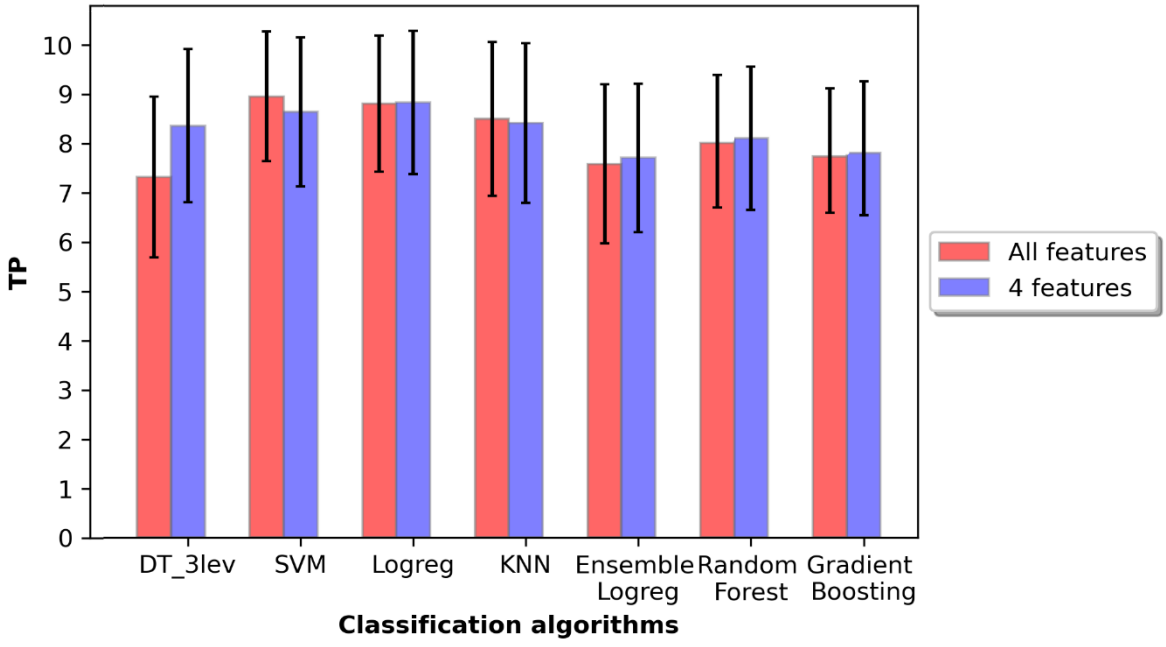


Figure 30. True positives for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

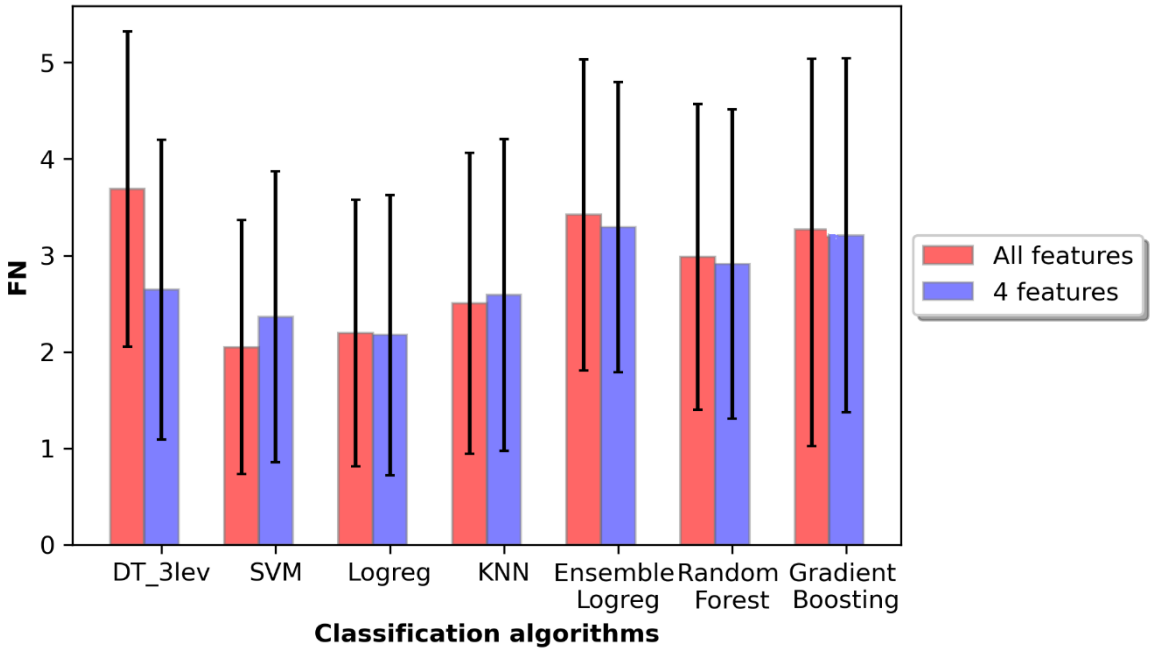


Figure 31. False negatives for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

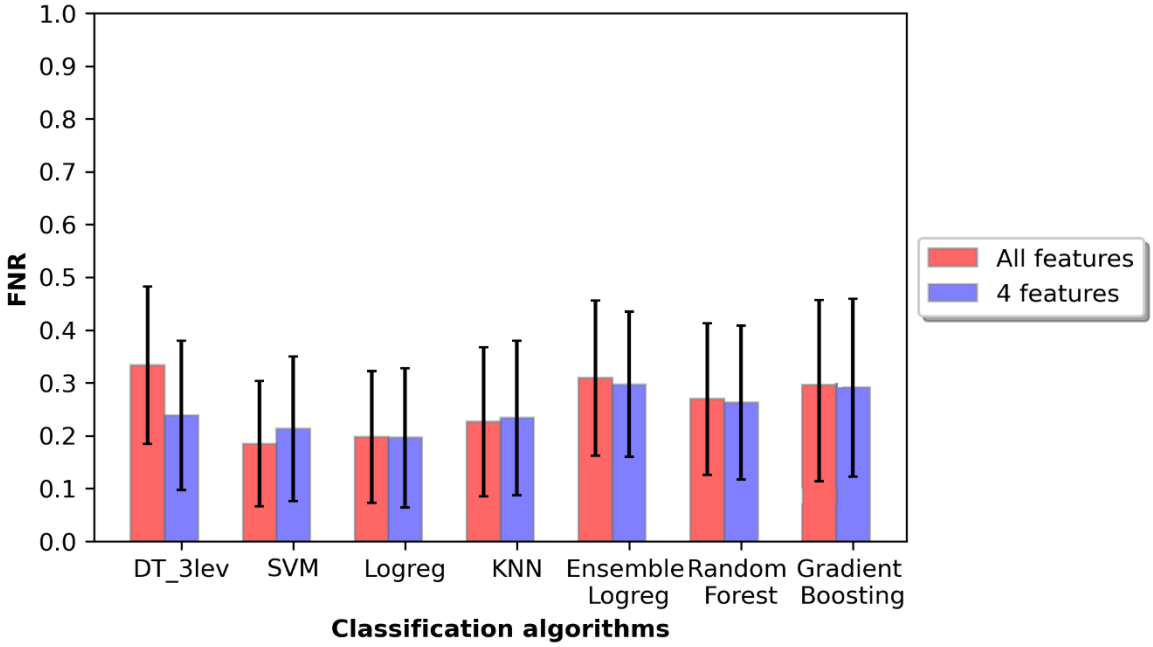


Figure 32. False negative rates for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

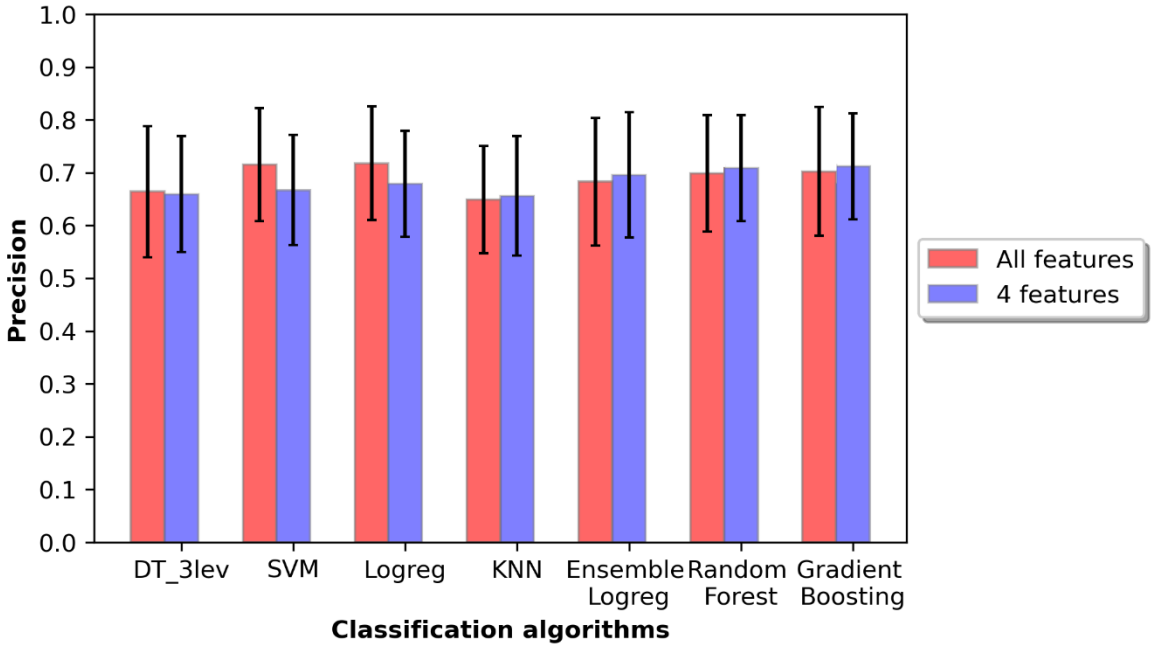


Figure 33. Precisions for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

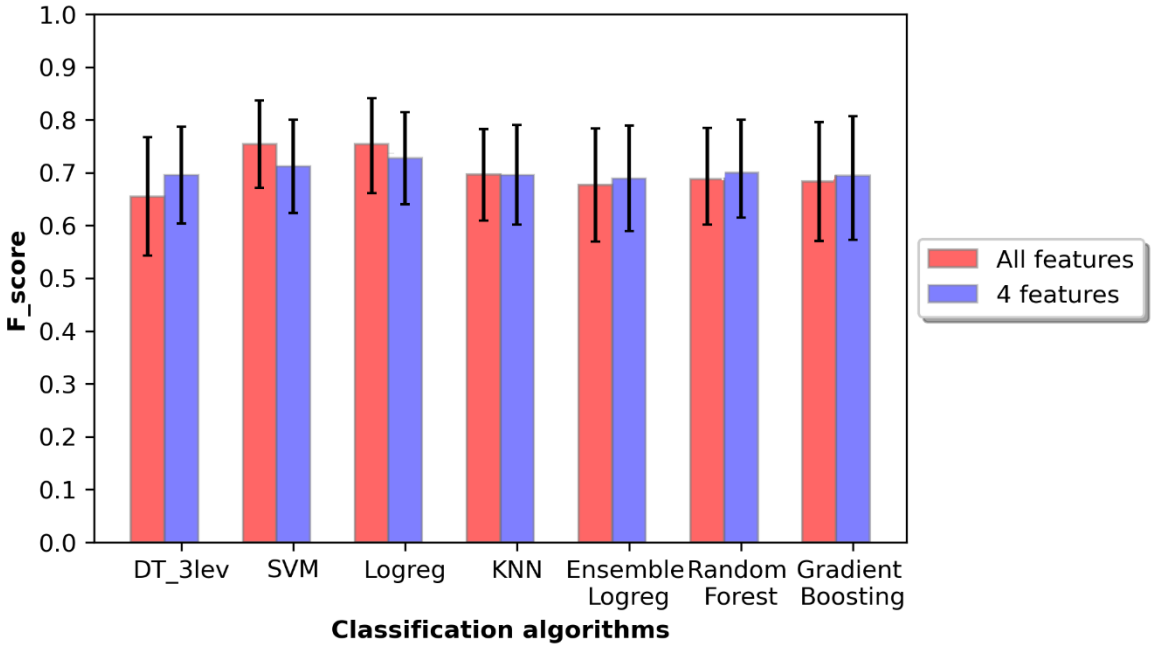


Figure 34. F-scores for the seven different machine learning approaches, considering the whole set of features and a subset of four features.

The figures presented in this section allow to have a visual comparison of the trend of the ten different metrics considered, for the seven different methods of supervised learning examined in this study. In particular, to optimize the visualization, and based on the fact that the metrics obtained analyzing the different feature sets were on average similar, it was decided to include only the values obtained when considering the entire set of then features and the set of 4 features (i.e., SRT, Age, #Correct and Avg_reaction_time).

Regardless ensemble logistic regression performance, both SVM and logistic regression were the approaches that showed the best training accuracy (i.e., fraction of correct predictions made by the classifier on the training set), as it can be noticed from Figure 25, with a value significantly higher with respect to the other methods, when considering the whole set of features (0.83 for both methods) and the subset of 4 features only (0.81 for SVM and 0.8 for logistic regression). The training accuracy of the ensemble logistic regression was actually much higher (0.95) with respect to the others (≈0.8); however, its test accuracy was much lower (0.78). This huge difference between test and training accuracy suggest the presence of overfitting, as the model was very good in learning past data (training set) however it was not flexible enough to generalize the fitted model to the training set and therefore the prediction error was quite high.

Looking at the other models instead, test accuracy is slightly lower than training accuracy, as training phase is carried out on labelled data while test phase is based on unknown data and therefore more misclassifications are in general expected, however, the difference between the two accuracies is minimal. The performance of SVM and logistic regressor were comparable and significantly higher with respect to the other techniques also with regard to the accuracy evaluated on the test set (see Figure 26); particularly, the test accuracies were distributed around 82% when considering all the features. Considering the models built from the 4 features, instead, all the machine learning approaches here considered behave in a similar way in terms of test accuracy, with an average value distributed around 0.78.

Even considering AUC (Figure 27), logistic regressor and SVM performed the best with a value of 0.91, whereas DTs were the worst with a much lower value, around 0.74, when considering the entire set of features. A considerably higher value (0.82), but still lower with respect to the other algorithms was found selecting a set of four features.

Specificity is a metric referring to the quote of individuals with no hearing impairment ('pass') correctly discriminated by the classifier. In terms of this metric (Figure 28), differences have been found considering the 2 set of features.

Including the whole set of features, all the techniques except for KNN presented a value of about 0.82. Specificity of KNN, instead, was equal to 0.77. Considering only 4 features, both Random forest and Gradient boosting presented a value of 0.83, which is higher with respect to the specificities found for the other methods ($\simeq 0.79$).

The sensitivity, shown in Figure 29, explains the percentual amount of hearing loss ('fail') correctly individuated. Regarding this metric, SVM and logistic regressor again performed hugely better than the other methods in terms of both set of features considered (whole set of features: 0.81 and 0.8; set of 4 features: 0.78 and 0.8; for SVM and logistic regressor, respectively). Also KNN performance was reasonably high, yielding a sensitivity of 0.77 for the entire set of features and equal to 0.76 considering 4 features only. The worst value (0.66) was found for the DT considering the whole set of features. Conversely, the DT based on 4 features only, presented a higher value of sensitivity (0.76), suggesting that the selection of a subset of features may create a more efficient set of splitting rules for the identification of hearing loss.

Precision regards the number of subjects actually presenting hearing loss, out of all the subjects that were predicted as having hearing loss. SVM and logistic regression performed the best considering the whole set of features, with a value of about 0.72, whereas no significant differences among the different machine learning approaches were found considering only 4 features.

True positive represents the amount of 'fail' correctly predicted. Ideally, we want this value to reach the number of 'fail' records present in the test set (i.e.,11). The best results were achieved using SVM and logistic regression, both when considering all the features (i.e., TP equal to 8.95 and 8.8), and the reduced set of 4 features (i.e., TP equal to 8.64 and 8.83).

False negatives instead represent the number of records that were actually 'fail' but has been classified as 'pass'. This metric needs to be minimized; indeed, a perfect classifier would have zero false negatives as all the subjects with hearing impairment are correctly recognized. A test presenting high sensitivity, in turn will provide a lower number of false negatives. The considerations regarding the best models in terms of FN coincide with what has been said for true positives, as the sum of TP and FP coincides with the number of

subjects having hearing loss, present in the test set. Indeed, the trend of the FN followed that of TP, with a better performance considering SVM (2.05) and logistic regression (2.2) while the worst values were found with the DT concerning the entire set of features (3.69).

False negative rate (miss rate) is another metric that takes into account both TP and FN and represents the probability that an individual with hearing loss will be missed out by the classifier. This probability needs to be minimized; lower values were again found for SVM and logistic regression, with a value of about 0.2, almost independent on the number of features considered, and higher values were obtained for DTs, with a certain difference between the 2 sets of features (i.e., 0.34 when considering all the features and 0.24 when considering 4 features).

When dealing with unbalanced classes, as in this case (101 pass and 55 fail), it should be also useful to calculate F-score which is a combination of precision and sensitivity that can be used to evaluate the performance of a classification model. Also for this last metric, SVM and logistic regression exhibited the best values (0.75), DT presented the lowest value (0.65) and the other techniques used provided results in between ($\simeq 0.7$), when considering the full set of features. Slightly lower values, with less differences among different classification methods were found for the reduced set of four values.

5. Discussion and conclusions

5.1 Distribution of variables according to ear tested, gender and age of the participants

The dataset has been analyzed in order to evaluate the possible influence of sex, ear tested (left/right) and age on the collected data, so to understand if the test main outcome would depend on these factors and if some sort of prevalence was present.

Hearing problems may not affect both ears in the same way, so it may be possible that, testing single ears, differences in the values for left and right ears in the presence of an unbalanced sample are obtained. As it is reported in the previous chapter (Table 2), the main results achieved for the right and left ears were similar, except for the average reaction time. This may be due to the fact that the participants who performed the test on the left ear were on average slightly older (about 5 years older) and perhaps may be slightly less reactive in replying to the proposed stimulus. Nevertheless, it can be affirmed that the test results are independent on the ear examined.

It was then evaluated whether gender could play a role in the test results. The starting sample was unbalanced, with almost twice as many women as men. However, no significant differences were found for any of the variables considered. Although it has been shown that women typically show less ability to hear low frequencies as they age, whereas hearing thresholds strongly increase in the high frequencies for men (Pearson, et al., 1995) (von Gablenz, Hoffmann, & Holube, 2020), the average hearing thresholds obtained after PTA assessment seemed not to be affected in this study, showing no evidence of a prevalence of hearing impairment due to the gender.

These initial analyses showed that the sample was immune to possible bias in the test outcomes due to the sex of the participant or the ear tested, in contrast, a significant effect of age was found, with higher age associated in worse test outcomes. In the present study, as the age range considered was very wide (20-89 years), three groups were considered: Young (i.e., age: 20 - 25 years), Adults (i.e., age: 25 - 60 years) and Elderly (i.e., age $\geq$ 60 years). Hearing loss is very often related to aging, presbycusis indeed is the hearing loss due to gradual changes and deteriorations in the auditory system (e.g. hair cell loss or structural stiffening) and in the nervous system (e.g. reduction in number of neurons and synapses in the auditory centers) with age. Indeed, by collecting a sample of unscreened individuals, it

is more likely that a person who has a hearing impairment but is not yet aware of it, is an adult or elderly person, while it is much less likely that a young person undergoing random testing has hearing loss. One of the main symptoms related to hearing loss with age, besides the elevation of the hearing thresholds, is the difficulty in understanding conversations in the presence of a background noise. Thus, speech-in-noise tests are a suitable tool to highlight the presence of age-related hearing loss.

Worse performances of speech recognition with age in individuals who took a SNT have been already demonstrated by comparing a sample of people over 60 years and a sample under 25 years (Heidari, Moossavi, Yadegari, Bakhshi, & Ahadi, 2018). The same conclusions, considering the newly developed SNT and 3 age-related groups based on the same cutoffs were reached in this study. One of the objectives of this first analysis was to highlight if this age-related deterioration could be spotted, not only in the older age group, but also in a group with subjects of intermediate age. Almost all the features analyzed, showed relevant age-related differences: young subjects revealed excellent speech recognition abilities in noise; adults showed moderate performance whereas elderlies reported poorer capabilities to recognize speech in presence of a background noise.

As previously cited, the speech reception threshold increased as age increased, since a worsening of speech recognition abilities in noise is one of the symptoms of age-related hearing loss, indeed, SRT increased from a mean value of -15.86 dB SNR in the Young group, to -12.86 dB SNR in the Adults group and finally increased again to -6 dB SNR in the Elderly group. Besides, also a significant increase in the average pure-tone threshold has been noticed, with an average PTA value of 0 dB HL in the Young group, of 20 dB HL in the Adults group and of 31 dB HL in the Elderly group, respectively, reflecting the fact that an increase in the hearing threshold is part of the gradual degeneration of hearing capabilities which typically becomes an effective problem from 60-65 years old and progressively worsen (Purnami, Mulyaningsih, Ahadiah, Utomo, & Smith, 2020).

In comparison with younger subjects, older adults gained higher scores in the HHIE-S questionnaire, stating that an increasing self-perceived hearing loss reflects the actual presence of a worsening in the hearing abilities. However, no difference has been noticed between adults and elderly, supporting the theory that, even if the actual performance in the test execution and hearing thresholds worsen from the Adults to the Elderly group, elderlies

tend to underestimate the perception of first clues of hearing loss as a deterioration in the ability to hear is considered as a natural age-related process and not an handicap.
As expected, the number of trials, the number of correct responses and the percentage of correct responses decreased with age as aged people performed worse in the test.
Conversely, the total test duration seemed not to be dependent on the age, due to a compensation effect; indeed young users required an higher number of stimuli before ending the test, but they responded quickly, on the other hand older subject replied to a lower number of stimuli, but each time they spent more time selecting the response.

5.2 Correlation between test variables

Several studies have analyzed the correlation of SRT extracted by speech-in-noise tests and the average pure-tone threshold of the 4 central frequencies assessed during audiometry (500, 1000, 2000 and 4000 Hz). SRTs derived from the newly developed SNT test showed a moderate correlation with the average pure-tone hearing thresholds (r=0.66). This finding is comparable with the correlation coefficient obtained for other online tests, namely Earcheck (r=0.66) and Occupational Earcheck (r=0.69) but it is slightly lower with respect to other SNTs including the National Hearing test (r=0.72) and the Dutch sentence SRT (r=0.82) (Leensen, de Laat, & Dreschler, 2011).
Thus, other attributes extracted from the test can be reasonably considered to improve the ability of the latter to distinguish between subject having no hearing loss and subjects with mild or higher degree of hearing loss.
Age was the feature which presented the strongest correlation with the average pure-tone thresholds (r=0.76), however also other variables such as the number of correct responses and the average reaction time were quite correlated with the screening result (r= -0.62 and 0.53, respectively), suggesting that it might be worth considering them as attributes for the classification task.
Regarding multicollinearity among classification attributes, the chosen test volume showed no significant correlation with the other variables, except for a weak negative correlation with the average reaction time. SRT, instead, showed a strong correlation (r=0.83) only with the percentage of correct responses and moderate correlations with almost all the other variables were observed.

The scatterplots for each couple of variables as a function of the two WHO criteria investigated (Figure 15 and Figure 16, respectively for criterion 1 and criterion 2) were analyzed so to highlight the presence of some data aggregation by target class. Variables that allow to distinguish quite well 'pass' from 'fail' may be useful attributes for classification.
Indeed, considering some couples of features, data points related to individuals with hearing loss ('fail') were distributed in closer proximity to one another (e.g. in correspondence of higher SRTs and older age, older age and higher number of trials, higher average reaction time and higher SRT values…) whereas for other couple of features (i.e., 'Volume', 'Total_test_time', and 'Score'), 'fail' records had a wider distribution, presenting an higher variability and were more difficult to be separated from the 'pass' ones. Indeed, a classifier based on those values would have a lower capability to provide the correct screening result.

5.3 Relationship between SRT and WHO criteria

The relationship between the speech reception threshold and PTA was investigated to assess if the SNT test main outcome alone could predict the degree of hearing impairment as defined by the WHO criteria starting from pure-tone-audiometry outcomes.
Since screening examinations target individuals searching for an early indication about their hearing conditions, only criterion for 'slight/mild' and 'moderate' hearing loss were addressed whereas 'severe' and 'profound' hearing loss detection was considered out of scope .
GLMs evaluating SRT against the binary target, showed a significant contribution of SRT for both mild (criterion 1) and moderate (criterion 2) hearing loss detection criteria. These findings confirm that SRT may be a useful attribute to be used in the classification task. Even if SRT allowed to distinguish pretty well 'pass' from 'fail', the distributions of the two classes partially overlap, as it can be noticed in the diagonal of Figure 15, therefore a trade-off between sensitivity and specificity of the classification would be necessary since there would be some misclassifications.
ROC curves were built for the 2 criteria by progressively varying the discrimination SRT and by selecting the best candidate SRT cut-off as the value closest to the ideal point (i.e. top left angle of the curve). Examined individuals with SRT higher than the cut-off fail the screening test (presence of hearing loss) whereas subjects with SRT below the cut-off value successfully pass the test. Previously analyses on a smaller population (98 subjects),

considering SRT as the only attribute for the classification into 'pass' (i.e., subject having hearing thresholds lower or equal to 25 dB HL at 1, 2, 4 kHz) and 'fail' (i.e., subjects having hearing thresholds higher than 25 dB HL at the same frequencies) exhibited a test accuracy of 0.82 and an AUC equal to 0.84. A specificity of 0.9 and a sensitivity of 0.7 were found selecting $SRT_{cut\text{-}off}$ equal to -8 dB SNR while a specificity equal to 0.81 and a sensitivity of 0.77 were found setting $SRT_{cut\text{-}off}$ to -10 dB SNR (Paglialonga, et al., 2020).

The discrimination capabilities here achieved for criterion 1, on 156 observations, by setting a $SRT_{cut\text{-}off}$ of -8.75 dB SNR, appeared to be fully comparable to the previous findings on a narrower sample, with an AUC of 0.83, a specificity of 0.84, sensitivity of 0.76 and test accuracy of 0.81.

Criterion 2 instead, setting a cut-off at -6 dB SNR led to better values in terms of AUC (0.89), sensitivity (0.89) and accuracy (0.83), however the specificity found was slightly lower (0.81). Comparing the observed performance for criterion 1 and criterion 2, the latter showed better values in terms of AUC, sensitivity, and accuracy and slightly lower specificity.

The Speech Understanding in Noise test (Paglialonga, Tognola, & Grandori, 2014), which is a SNT based on the same speech material (meaningless VCVs) and recognition task (3AFC), but uses a fixed-levels testing procedure, exhibited a specificity and a sensitivity of about 0.84 (83.9% and 83.8%) against the WHO criterion for moderate hearing impairment (criterion 2). These findings match the classification performance of the SNT here considered, despite so, further assessments need to be done in presence of a population including a higher number of subjects suffering from moderate hearing loss, as only a few subjects in the tested population (i.e., 18) belonged to the 'fail' class.

The results obtained in terms of prediction accuracy for both criteria, considering SRT as the unique predictor for hearing loss, are moderate and may be explained by the fact that the independent and the dependent variables refer to two different phenomena, namely speech recognition for what concerns SRT and pure-tone sensitivity regarding WHO criteria based on PTA.

As it could be expected, the cut-off value determined for criterion 2 is higher with respect to the one assessed for criterion 1, as it aimed at the discrimination of a more severe hearing loss. SRT alone is not sufficient to explain the dependent variable, indeed the goodness-of-fit metric of the model, represented by the adjusted R squared, resulted to be only moderately

high and much lower for criterion 2 (0.23) with respect to criterion 1 (0.34). Therefore, also the other variables or at least a subset of them must be used to improve the classification.
As a first step, age was used in addition to SRT to predict the screening result as the increase of pure-tone thresholds is one of the factors associated with ageing (von Gablenz, Hoffmann, & Holube, 2020). Age, together with SRT resulted to be a significant predictor for criterion 1 whereas was not significant for criterion 2. These results could be supported by the fact that more severe hearing loss may be associated to different causes less related to age.
Finally, the interaction between SRT and age was found to be meaningless for both criteria, meaning that, the association between SRT and the probability to fail the test is independent from age.

5.4 Relationship between other test variables and WHO criteria

Single-predictor GLMs evaluating the other 9 attributes against the binary PTA showed that, besides the volume and the test duration, all the other variables contributed significantly to the prediction for both WHO criteria for hearing impairment. However, the goodness-of-fit of the models was quite low, and reached its maximum value considering age as predictor for criterion 1 (0.4) and considering the correct number of responses as predictor for criterion 2 (0.26).
The implementation of GLMs that included the SRT and each one of the other features as predictors, allowed to explain an higher amount of variance of the dependent variable, for both criteria, with a major increase in R^2_{adj} when considering features like 'Age', 'Score', '#Trials' and '#Correct'. Concerning the other models, the addition of a new predictor did not cause any significant increase in the ability to explain the dependent variable, with respect to SRT alone. The maximum R^2_{adj} values were achieved considering age and SRT as predictors for criterion 1 (0.44) and considering the score of the HHIE-S questionnaire and SRT as predictors for criterion 2 (0.33).
The volume and the total test duration, as it could be expected by the correlation analysis, were not able to explain the target variable and therefore their presence might not bring important contributions to classification of hearing impairment.
The lack of importance of the selected volume in the screening result may be related to the fact that most participants considered the default volume suitable for test execution and therefore did not adjust its value. The total test duration may also be a negligible feature for

the classification task as it does not vary considerably with changes in the hearing thresholds of the tested individuals. This is mainly due to a compensatory action, as subjects who perform well in the test have to go through a higher number of trials (i.e., to reach a lower SRT) before reaching the end of the test (i.e. 12 reversals), but they tend to respond faster at each trial because they are less hesitant in the correct detection of the VCV masked by noise. Conversely, people with worse speech recognition performance have to deal with a lower number of trials, but they can took on average the same total time as good performers, because of their longer response time due to a higher hesitation in giving their answers. As a result, subjects tend to take about the same time to perform the test, almost independently on test outcomes.

Regarding the interaction terms, only the interaction between SRT and the test duration has been found to be significant for criterion 1. A longer duration of the test may be due to a better hearing, which means that the subject listens to more stimuli before finishing the test and has a lower SRT, resulting in a successful completion of the screening procedure.

However, even if the subject struggles to distinguish stimuli and therefore responds with indecision, although answering fewer questions, the time spent on the test will be high. In this case, the SRT will be worse (higher) and for this reason the screening result will be a failure. Therefore, the presence of a significant interaction means that the effect of the test duration on the result of the screening is different for different values of SRT.

5.5 Feature selection

Besides using the full set of features, seven different subsets have been evaluated as a selection of a subset of features may reduce overfitting by removing redundant and unimportant data, and also speed up the training process that could require a certain amount of time with a larger sample of data. Features which result to be irrelevant may not only not improve the performance but also negatively affect the classification performance, reducing accuracy.

Several studies validated the HHIE-S against pure-tone audiometry, however different values of sensitivity and specificity were found, mostly depending on the age of the groups considered, the numerosity and the type (i.e., level of education achieved) of the population analyzed, and the definition of hearing loss used. Though, HHIE-S in general shows higher

sensitivity and specificity when dealing with severe hearing loss detection, with respect to mild and moderate hearing loss. (Servidoni & De Oliveira Conterno, 2018). In particular, a validation study of the HHIE-S for the Indian rural elderly population on 175 subjects (Deepthi & Kasthuri, 2012) reports a sensitivity of 76.2% and a specificity of 87.7% when dealing with severe hearing loss, whereas a consistently lower value of sensitivity (26.2%) was found for weaker hearing impairments.

The Blue Mountains Hearing Study (Sindhusake, et al., 2001) instead obtained a sensitivity of 58% and specificity of 85% when measuring mild hearing loss and a better sensitivity (100%) and slightly lower specificity (70%) when considering marked hearing loss, in a population of 2015 subjects. These results may suggest that HHIE-S may be more effective in the discrimination of moderate and severe hearing loss, that are out of the scope of the speech-in-noise test analyzed in this study, but may under-detect mild hearing impairment.

The two studies just reported also highlighted a decrease in self-perception of the hearing handicap with age. Once considered the degree of hearing impairment, indeed the HHIE-S has best performance in younger subjects with respect to older ones (Wiley, Cruickshankst, Nondahlt, & Tweed, 2000), perhaps because young people are more aware of hearing loss as a possible issue in terms of social and emotional limitations to participation, whereas older subjects consider a degradation of the hearing abilities as a natural consequence of aging.

Besides, elderly patients are less exposed to noise with respect to younger ones and may perceive less to have hearing problems (Rosdina, Leelavathi, & Azimah, 2011).

Indeed, it is more likely for an elderly individual to have hearing loss, but not report it. These underestimations of hearing impairment could bring to bias and an increase of variance in a classification task. Moreover, the level of education may affect the ability to fill out the questionnaire and people with a low level of education (i.e., elementary school level) may require additional explanations and support (Purnami, Mulyaningsih, Ahadiah, Utomo, & Smith, 2020), preventing the development of a mass-screening procedure.

The test submission may last a few minutes, at most 10 (Servidoni & De Oliveira Conterno, 2018), and this can impact a lot on the total time of the screening procedure, going to nullify what is the advantage in terms of time of execution of the new SNT, without significantly improving the performance of the classification. Conversely, the removal of the score of the questionnaire from the attributes of the classifiers doesn’t imply a considerable loss in terms of performance as it can be seen comparing the values obtained for models considering the

whole set of features and the ones obtained considering only nine features (removing in turn the raw score and the degree of hearing impairment obtained from the questionnaire score) or eight features (removing both variables related to the questionnaire).

In the context of feature selection, the analysis of the correlation matrix may be useful to individuate a subset of features by understanding how the different attributes are related each other and to the target variable (i.e., the screening result). First, features that were not significantly correlated with the target variable, namely volume and the total duration of the test, were removed as they added no useful information in the target prediction. The fact that there were no significant changes in the trend of predictions considering the group with all the features and only six features (removing 'Volume' and 'Total_test_time' in addition to 'Score') justifies the removal of these two variables from the classifier in this preliminary evaluations , as they may not affect the prediction of screening results.

Features that were highly correlated each other (considered as having absolute value of correlation coefficient higher than 0.8), resulted to have quite the same effect on the target variable, and therefore one of them could be removed without affecting the overall performance as it bring redundant information. In addition, the presence of multicollinearity between attributes may lead to erratic predictions when dealing with some classifiers such as logistic regressors. The feature which had the highest correlation were '#Trials' and '#Correct' (r=0.99), among them, the feature providing the highest absolute correlation coefficient with the target was '#Correct', therefore exploiting a greedy elimination approach, the number of trials, which had the lower correlation with the PTA outcome was removed from the set of features. In the second place, '%Correct' and SRT were highly negatively correlated (r=-0.83), among them, the SRT had the highest correlation with the PTA and therefore was maintained, whereas '%Correct' was discarded.

Finally, the 4 candidate features were: SRT, age, number of correct responses and average reaction time.

5.6 Evaluation of classification outcomes

The selection of training and test partitions in datasets with reduced size, as the one here considered, may potentially introduce variability in the classification, which was addressed by running 1000 iterations of the models and considering the average performance. The

standard deviations were relatively low (i.e., smaller than 0.1) for parameters like the accuracies and AUC, they reached values around 0.1 for F-measure and precision and slightly higher values for FNR, precision, and sensitivity (e.g. 0.17, 0.12 and 0.17 considering gradient boosting). The average performance of the models was similar, both considering different classification algorithms and different sets of features. Indeed, the limited differences in the classification performance using different sets of features may be due to the intrinsic variability of the models built starting from different partitions of the original dataset in training and test sets. Despite so, on average, the different machine learning approaches perform quite similar to each other regarding the metrics evaluated. Besides the performance indicators, also the general drawbacks and limitations of the algorithms here considered must be discussed.

Decision trees gave the more interpretable results, with simple and justifiable decision rules and fast predictions that cope very well with the requirements of the smartphone app to get a real-time screening result. They are robust to both outliers and missing data, thus they do not require particular data pre-processing and they are not affected by collinearity of the attributes, however they are more prone to overfitting and they are models with high variance, meaning that the performance is very dependent on the training data and even small differences in the training data may bring to different classification results.

SVMs instead are less prone to overfitting and perform very well even with a large number of attributes, however they are in general slower with respect to DTs, they require features scaling to optimize the kernels computation and they provide less interpretable results.

KNNs do not need to tune parameters besides the number of neighbors and the distance metric, and they are robust to the presence of non- significant features, however as SVM they require feature normalization. Even KNN approach gets significantly slower and complex when dealing with a much bigger dataset, moreover it is very sensitive to the presence of correlated features (multicollinearity).

Logistic regression algorithms are quite fast for both training and predictions; they also provide the probability of the observation to belong to the class, however they are usually not flexible enough in presence of non-linear separations. While SVM can support both linear and non-linear separation problems thanks to the use of kernels, logistic regression works well only when data are linearly separable which may not be the case in most of the real-life problems bringing to possible underfitting (because a linear dependency on the data

is assumed) and a decrease in accuracy, moreover its performance may be influenced by the presence of outliers and correlation between independent variables.
Ensemble logistic regression instead combine the result of the previous methods and use them to train a meta-model, of course the obtained result are less interpretable, the model complexity is higher and results depend on the kind of weak learner chosen and the training data, especially when dealing with small datasets.
Random forests reduce the variance with respect to individual classification trees despite a loss in interpretability. They are usually less prone to overfitting when compared with single trees, however, increasing the number of trees in the forest may also slow down the time required for prediction with respect to other methods. Random forest can also return information about the relative importance of each feature with respect to the classification, providing a clue about which features to select
Gradient Boosting is a powerful method, it is flexible and does not need any data pre-processing. However, it is highly sensitive to outliers, it usually requires more time to train and a more complex hyper-parameters tuning.

As it is well known, there is not a perfect algorithm which is able to optimally solve each kind of classification problem and outperforms the others in all kind of situations ('no free lunches' concept) however, depending on the task to execute, one algorithm may behave better than the others.

Seven different supervised learning techniques were compared. Regarding the classification task here discussed (i.e, screening tool for hearing loss detection), all the algorithms provide consistent results.
SVM and logistic regression seemed to provide slightly more promising results, Gradient boosting and Random forest with a reduced number of trees (i.e., 10) performed well too, however the prediction time of the latter would increase a lot by increasing the forest size.
Further analysis on a much larger sample are required to reduce the variability and stabilize the performance.
The SVM model considering the whole set of features reported satisfactory mean performance over 1000 iteration, indeed the accuracy on the test set reached 0.82 , the area under the ROC curve was equal to 0.91, the specificity was equal to 0.82 and the sensitivity

was equal to 0.81. The values obtained considering 4 features were on average one percentage point lower.

The logistic regressor, that also showed encouraging results, reached an accuracy on the test set equal to 0.82, an area under the ROC curve equal to 0.91, a specificity equal to 0.82 and a sensitivity equal to 0.8. Even for this model values obtained considering 4 features were on average slightly lower or comparable.

Random forest and gradient boosting showed the same specificity of SVM and logistic regression, and provide also good values in terms of AUC (0.87) and accuracy on the training set (0.79 and 0.78 respectively) However, their ability to detect subjects with hearing loss, represented by sensitivity, appears to be too low (0.73 and 0.7) and at the moment prevent them from being used as screening tool. Random forests showed promising results for the purposes of the study, as the main future goal will be to create a smartphone app able to provide quick and real-time screening results, the number of trees in the forest should be limited to avoid excessive prediction times.

KNN presented an AUC equal to 0.85 and a specificity of 0.77 which resulted to be lower than that of the previous methods. However, it yielded a test accuracy comparable to the one found for Gradient boosting (i.e., 0.77) and a better sensitivity (0.77) with respect to Random Forest and Gradient boosting.

The ensemble logistic regressor underwent overfitting, as it can be seen from the huge difference between training and test accuracy (0.95 and 0.78, respectively), and for this reason it was not further addressed in this preliminary evaluation as it may likely produce significant prediction errors on new records. A finer tuning of the hyperparameters, combined with the choice of different models could be further implemented in future to prevent this problem.

Lastly, the classification performance of DTs in this preliminary study seemed to be the lowest, especially in terms of AUC (0.74). The AUC obtained with DTs using a reduced set of 4 features is higher (0.82), but still lower compared to other classification algorithms.

The results obtained considering the prediction of the presence of a hearing loss exclusively from the speech reception threshold, selecting a $SRT_{cut\text{-}off}$ of -8.875 dB SNR and reported in Table 5, were lower in terms of sensitivity and AUC as compared to SVM and logistic regression performance, but absolutely equivalent in terms of specificity and accuracy on the test set.

Furthermore, the capabilities of the classifiers here introduced to detect subjects with hearing impairment are compatible with the results found in the literature for the application of other speech-in-noise tests based on similar approaches, to be used for screening purposes.

The performance of the new SNT is slightly worse than the already mentioned SUN test, but the use of different WHO criteria to classify hearing loss (respectively 25 dB HL and 40 dB HL) in the two tests certainly has an impact on the discrepancy between the results.

A multi-centric study carried out by Leensen, de Laat and Dreschker in 2011 (Leensen, de Laat, & Dreschler, 2011) compared results of the Dutch National Hearing Test, Earcheck and Occupational Earcheck (OEC) with pure-tone thresholds to assess their performance in NIHL detection. Earcheck showed a high specificity equal to 0.90 and a much lower sensitivity (0.51) meaning that normal-hearing subject were correctly classified in most of the case whereas NIHL detection were missed in almost half of the subjects actually presenting hearing impairment. Occupational Earcheck, which is a slightly modified version of the previous one, showed instead a very low specificity (0.49) and a much higher sensitivity (0.92).

Both logistic regression and SVM models that used the entire set of features collected demonstrated on average a better ability to identify subjects with hearing loss (higher sensitivity), but a lower ability to classify healthy subjects (lower specificity) when compared to Earcheck. On the contrary, the classifiers here implemented, showed a higher specificity and a lower sensitivity in recognizing the presence of hearing impairment than OEC.

When dealing with the evaluation of the performance of a screening test, both sensitivity and specificity values may be highly influenced by the speech material, noise selection, SRT definition, selected cut-off and of course the selected dependent variable i.e. frequency to be used to average the hearing thresholds. Indeed, further studies of the previously cited authors (Leensen, De Laat, Snik, & Dreschler, 2011) showed that manipulation of the masking noise and in particular the use of a low pass filtered masking noise, could improve both sensitivity and specificity in a controlled environment. More recently, a cross-sectional study proposed by Rashid et al. (Sheikh Rashid & Dreschler, 2018) highlighted a benefit of automatic conditional rescreening in the use of Occupational Earcheck as a screening test for high frequency hearing loss (HFHL), on a population of 80 subjects habitually exposed to noise in their workplace. An improvement in test specificity has been spotted after rescreening,

passing from a value of 0.63 to a 0.93, whereas sensitivity on the whole population instead slightly decreased from 0.65 to 0.59. The older population seemed to benefit more from the presence of a second trial, with specificity increasing from 0.46 to 0.92.

The National Hearing Test, a digits-in-noise screening test delivered by telephone, based on the presentation of stimuli composed by a sequence of three digits with a background noise, was found to have a specificity of 0.93 and a sensitivity of 0.91 (Smits, Kapteyn, & and Houtgast, 2004). Leensen et al. obtained instead a comparable value of specificity (0.94) but a much lower value of sensitivity (0.55). The digits-in-noise test has been adapted and validated in several countries, among those, the US version yielded a specificity of 0.83 and a sensitivity of 0.80 using pure-tone threshold higher than 20 dB as discrimination criterion to identify individuals with hearing impairment and a $SRT_{cut\text{-}off}$ of -5.7 dB (Watson, Kidd, Miller, Smits, & Humes, 2012). A computer-based adaptation of the US-DIN (Folmer, et al., 2017) has been evaluated against the presence of hearing loss based on criterion 1 showing an AUC equal to 0.95 considering both ears. Findings were also similar for each ear analyzed separately.

Both the logistic regression and SVM models based on the new SNT provided similar values compared to the American version of the digit-in-noise test in terms of accuracy, specificity and sensitivity, and slightly lower values, but still consistent, in regards to AUC.

The classification performance obtained using a set of features extracted from the new cross-cluster adaptive speech-in-noise test, are absolutely in line with the capability of some of the most popular SNTs to detect the presence of hearing loss.

5.7 Innovations

The present study introduces some innovative aspects. In several occasions the effect of aging in speech perception in noisy environments has been analyzed, however, the studies found in the literature are mainly based on the distinction into two groups, a group of younger adults and a group of older adults. For example, Schoof et al. (Schoof & Rosen, 2016) investigated the increasing of speech reception thresholds with age in a group of 19 younger subjects (i.e., 19-29 years) and 19 older subjects (i.e., 60-72 years) in presence of steady-state noise, amplitude-modulated noise and two-talker babble. Vermeire et al. (Vermeire, Knoop, De Sloovere, Bosch, & van den Noort, 2019), instead, measured the SRT, as

extracted by the Leuven Intelligibility Sentence Test, to evaluate the effect of age in sentences recognition in noise in a group of 27 young adults (i.e., 19.1-25 years) and a group of 33 older adults (i.e., 60.4-82.7 years). Finally, Heidari et al. (Heidari, Moossavi, Yadegari, Bakhshi, & Ahadi, 2018) demonstrated the deleterious impact of age on speech recognition in noise in a group of 32 elderly people (i.e., over 60 years) with respect to a group of 32 young people (i.e., 18-25 years). The present study, by introducing a third intermediate age category (i.e., 25-60 years) between young adults (i.e., 20-25 years) and older adults (i.e., over 60 years), allows to emphasize even more clearly the gradual and progressive loss of verbal recognition in a wider age range. In addition, the use of a test based on meaningless stimuli (i.e., VCVs), instead of sentences, has allowed to overcome possible bias due to different levels of education or native languages in the considered population.

Furthermore, researches aimed at evaluating the performance of speech-in-noise tests in the identification of the presence of hearing loss, normally use only the primary outcome of the test, i.e. the SRT, as the unique variable from which to predict the result of the screening.

In this thesis work, instead, a multivariable approach has been used, estimating the prediction of the screening result according to the WHO criteria for hearing impairment from a series of features, not only those directly related to the execution of the SNT test, but also the score obtained in the ten questions of the HHIE-S questionnaire and the age of the individual.

A satisfactory classification performance, comparable with results found in literature for other SNTs, was obtained in the detection of hearing loss according to the WHO criterion for mild hearing impairment (i.e., PTA> 25 dB HL). Preliminary evaluations were also carried out using SRT alone to detect hearing loss according to the WHO criterion for moderate hearing loss (i.e., PTA> 40 dB HL), leading to promising results that need to be further validated.

5.8 Limitations and further research

The present thesis work provided quite satisfactory results, but it has some limitations.

The classification performance considering a sample of only 156 records was data-driven as running different times the same model brought to different values of the metrics, therefore an average performance over 1000 iterations was considered in order to provide a preliminary evaluation. First, the increasing of the sample size is needed to reduce the

variability of the classification process. Further studies on a larger population may allow to reduce this variability, by reducing the data-dependency, as both training and test partition would be more equally representative of the original dataset, when increasing the sample size. After doing so, optimized model definition and finer hyperparameters tuning could be achieved in order to further help towards the development of a smartphone application for adult hearing screening.

Collecting a dataset with a considerably larger size may also allow to analyze the performance of neural networks, not considered in this study as they require a great number of records to guarantee a good classification performance.

Another possible way to improve the abilities to detect hearing loss may be to evaluate other attributes to be added to the classifier, besides those here considered, such as risk factors like noise exposure, use of ototoxic drugs and pregressed medical history.

Additional research on a substantially larger population involving a higher number of subjects with hearing impairment, also with higher degrees of hearing loss (i.e., moderate, severe and profound) is a crucial step needed to validate the speech-in-noise test and the classification methods for both criterion 1 and criterion 2, also allowing a more precise definition of the SRT_{cutoff} value.

5.9 Conclusions

Politecnico di Milano, in collaboration with Consiglio Nazionale delle Ricerche, has recently developed a fast, automated and adaptive speech-in-noise test, which demonstrated to have high accuracy, test-retest repeatability and a reduced language dependency due to the use of meaningless VCVs as speech material. A preliminary evaluation of the viability of the implemented procedure as an adult hearing screening to promote early identification of hearing loss has been performed in this thesis work.

In a first stage, the impact of sex, ear tested and age on the distributions of the test outcomes was investigated, showing that there were no evident effects due to sex and ear tested, whereas a worsening in the test outcomes gradually occurs as age increases, reflecting poorer capabilities to recognize speech in a background noise. The decrease in the test outcomes do not happen drastically from young subjects to elderlies, but a progressive and significant worsening is also witnessed considering an intermediate group of adults.

Seven different classification techniques, including four of the most widespread and consolidated approaches (Decision Trees, Support Vector Machine, logistic regression, K-Nearest-Neighbors) and three ensemble methods (ensemble logistic regression, Random forest and Gradient boosting) were then addressed to analyze the feasibility of the novel speech-in-noise test as an adult hearing screening test for hearing loss identification according to the WHO criterion for mild hearing impairment (i.e. average pure-tone threshold > 25 dB HL). The analysis was carried out by using seven different sets of features suitably chosen starting from the examination of the correlation matrix and the strength of their relation with the target variable, together with their distribution according to the two classes ('pass' and 'fail') related to the screening result. The choice of the features was further confirmed by the analysis of single-predictor GLMs and multiple-predictor GLMs (i.e., including SRT and each of these features).

The investigated supervised learning methods reached moderate accuracies in detecting hearing loss, even considering a reduced number of features, demonstrating that the total duration of the test, the volume and the score of the HHIE-S questionnaire may not contribute significantly to the classification of ears into 'pass' and 'fail' class.

Among the evaluated classification approaches, SVM and logistic regression achieved the most promising results, fully comparable with those found in the literature for other well-known automated speech-in-noise tests.

A test developed to be used for hearing screening purposes should have high values for both sensitivity and specificity to correctly discriminate individuals presenting hearing loss from those who have normal hearing. Hence, the values of sensitivity obtained in this study should be improved, as, at the moment, about 20% of the examined subjects with hearing loss are erroneously classified as normal hearing.

Further research needs to be carried out to validate the speech-in-noise test, and the experimental procedure, with the final aim to implement a mobile app for adult hearing screening.

References

Agresti, A. (2013). *Categorical Data Analysis, Third Edition.*

Amplaid a137-a177 plus instructions manual. (n.d.). *[Manual]*.

Byrne, D., Dillon, H., & Tran, K. (1994). An international comparison of long-term average speech spectra. *The journal of the acoustical society of America. Vol. 96(4), 2108–2120*.

Cooke, M., Lecumberri, M. L., 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. Vol. 52 (11-12),954-967*.

Deepthi, R., & Kasthuri, A. (2012). Validation of the use of self-reported hearing loss and the Hearing Handicap Inventory for elderly among rural Indian elderly population. *Archives of Gerontology and Geriatrics*. *Vol. 55(3), 762-767*.

Folmer, R. L., Vachhani, J., McMillan, G. P., Watson, C., Kidd, G. R., & Feeney, M. P. (2017). Validation of a Computer-Administered Version of the Digits-in- Noise Test for Hearing Screening in the United States. *Journal of the American Academy of Audiology*. *Vol. 28(2), 161-169.*

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*. *Vol. 38 (4), 367-378.*

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 & Otology*. *Vol. 22(3), 134-139.*

Holder, J. T., Levin, L. M., & Gifford, R. H. (2018). Speech recognition in noise for adults with normal hearing: agenormative performance for AzBio, BKB-SIN, and QuickSIN. *Otology & Neurotology*. *Vol. 39 (10), e972-e978.*

Jansen, S., Luts, H., Dejonckere, P., van Wieringen, A., & Wouters, J. (2013). Efficient hearing screening in noise-exposed listeners using the digit triplet test. *Ear and hearing*. *Vol. 34(6), 773-778.*

Killion, M. C., & Niquette, P. A. (2000). What can the pure tone audiogram tell us about a patients. *The Hearing Journal. Vol. 53(3), 46-53*.

Kumar, U. (2018). *Applications of Machine Learning in Disease Pre-screening.*

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*. *Vol. 50(11), 823-834.*

Leensen, M. C., De Laat, J. A., Snik, A. F., & Dreschler, W. A. (2011). Speech-in-noise screening tests by internet, Part 2: Improving test sensitivity for noise-induced hearing loss. *International Journal of Audiology*. *Vol. 50(11), 835-848.*

Martens, M. K., Perenboom, R., Ploeg, C., & van der Dreschler, W. (2005). *Oorcheck: Analyse van gehoortesten voor jongeren over de periode april 2004 – oktober 2004. TNO-rapport.*

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models. 2nd Edition.*

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning. Secon Edition.*

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*. *Vol. 29(3S), 535-684.*

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*, *Vol. 52, 66-72.*

Park, K. V., Oh, K. H., Jeong, Y. J., Rhee, J., Han, M. S., Han, S. W., & Choi, J. (2020). Machine learning models for predicting hearing prognosis in unilateral idiopathic sudden sensorineural hearing loss. *Clinical and Experimental Otorhinolaryngology*. *Vol. 13(2),148-156.*

Pearson, J. D., Morrell, C. H., Gordon-Salant, S., Brant, L. J., Jeffrey Metter, E., Klein, L. L., & Fozard, J. L. (1995). Gender differences in a longitudinal study of age-associated hearing loss. *Journal of the Acoustical Society of America*. *Vol. 97(2), 1196-1205.*

Polo, E. M., & Zanet, M. (2018). Development and evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing. (Master's Thesis) Politecnico di Milano.

Polo, E. M., Zanet, M., Lenatti, M., van Waterschoot, T., Barbieri, R., & Paglialonga, A. (2020). *Development and Evaluation of a Novel Method for Adult Hearing Screening: Towards a Dedicated Smartphone App. 7th EAI International Conference on IoT Technologies for HealthCare (EAI HealthyIoT 2020).*

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 and Neck Surgery*.

Rocco, G. (2018). Design, implementation, and pilot testing of a language-independent speech intelligibility test. (Master's Thesis) Politecnico di Milano.

Rosdina, A., Leelavathi, M., & Azimah, M. A. (2011). Screening for hearing impairment among the elderly using hearing handicap inventory for the elderly-screening (HHIE-S). *The New Iraqi Journal of Medicine. Vol. 7(3), 68-72.*

Servidoni, A. B., & De Oliveira Conterno, L. (2018). Hearing loss in the elderly: Is the hearing handicap inventory for the elderly - Screening version effective in diagnosis when compared to the audiometric test? *International Archives of Otorhinolaryngology. Vol. 22(1), 1-8.*

Sheikh Rashid, M., & Dreschler, W. A. (2018). Accuracy of an internet-based speech-in-noise hearing screening test for high-frequency hearing loss: incorporating automatic conditional rescreening. *International Archives of Occupational and Environmental Health. Vol. 91(7), 877-885*.

Sindhusake, D., Mitchell, P., Smith, W., Golding, M., Golding, M., Newall, P. D., & Rubin, G. (2001). Validation of self-reported hearing loss. The Blue Mountains Hearing Study. *International Journal of Epidemiology 2001. Vol. 30(6),1371–1378.*

Smits, C., Kapteyn, T. S., & and Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology, Vol. 43 (1), 15-28.*

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. *The Journal of the Acoustical Society of America. Vol. 91(1), 421-437.*

Tomioka, K., Ikeda, H., Hanaie, K., Morikawa, M., Iwamoto, J., Okamoto, N., Kurumatani, N. (2013). The Hearing Handicap Inventory for Elderly-Screening (HHIE-S) versus a single question: Reliability, validity, and relations with quality of life measures in the elderly community, Japan. *Quality of Life Research. Vol. 22(5), 1151-1159.*

Townend, O., Nielsen, J. B., & Jesper, R. (2018). Real-life applications of machine learning in hearing aids. *Hearing Review. Vol. 25(4), 34-37.*

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*.

Van Eynde, C., Denys, S., Desloovere, C., Wouters, J., & Verhaert, N. (2016). Speech-in-noise testing as a marker for noise-induced hearing loss and tinnitus. *B-Ent, Vol. 12(26/1), 185-191*.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.*

von Gablenz, P., Hoffmann, E., & Holube, I. (2020). Gender-specific hearing loss in German adults aged 18 to 84 years compared to US-American and current European studies. *PLoS ONE. Vol. 15(4).*

Watson, C. S., Kidd, G. R., Miller, J. D., Smits, C., & Humes, L. E. (2012). Telephone Screening Tests for Functionally Impaired. *Journal of the American Academy of Audiology, Vol. 23(10), 757-767.*

Wiley, T. L., Cruickshankst, K. J., Nondahlt, D. M., & Tweed, T. S. (2000). Self-Reported Hearing Handicap and. *Journal of the American Academy of Audiology. Vol. 11, 67-75.*

World Health Organization. (2018). Addressing the rising prevalence of hearing loss. Geneva.

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. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. Berlin, Germany, July 23-27, 2019*

Zhao, Y., Li, J., Zhang, M., Lu, Y., Xie, H., Tian, Y., & Qiu, W. (2019). Machine Learning Models for the Hearing Impairment. *Ear and Hearing. Vol. 40(3), 690-699.*