MULTIVARIATE SCREENING AND FEATURE ENGINEERING FOR STATISTICAL DECISION MAKING IN DISEASE SEVERITY DIAGNOSTICS

Marta Campi
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
mcampi@pasteur.fr

Gareth W. Peters
Department of Statistics and Applied Probability
University of California Santa Barbara
Santa Barbara, CA 93106, United States
garethpeters@ucsb.edu

Perrine Morvan
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
pmorvan@pasteur.fr

Mareike Buhl[†]
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
mbuhl@pasteur.fr

Hung Thai-Van[†]
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
htaivan@pasteur.fr

May 21, 2025

ABSTRACT

Mapping multivariate continuous health measurements to discrete diagnostic or disease severity categories presents a persistent challenge in clinical practice, particularly with the increasing automation of diagnostics and the push toward standardized, evidence-based decision-making. We propose a three-stage statistical framework to address this challenge in a principled and generalizable manner. The methodology requires access to patient samples that include multivariate diagnostic test results alongside expert-assessed disease severity classifications, including the possibility of no disease. In the first stage, diagnostic tests are selected based on clinical focus, yielding a multivariate profile of physiological responses in the relevant diagnostic space. In the second stage, we transform these test responses into an abstract feature space through: (1) feature construction, (2) statistical evaluation and ranking of each feature's discriminatory power using hypothesis testing, and (3) bootstrapping to address class imbalance across disease severity categories and improve generalization. In the third stage, we learn a partitioning function from the feature space to discrete diagnostic or severity categories. This mapping can then be applied to new patients to support systematic, replicable disease classification. We illustrate this framework using audiological diagnostics, where hearing-loss severity categorization relies on combining pure-tone audiogram thresholds with speech-recognition scores. This domain, characterized by complex interdependencies between measurements, offers a compelling real-world test case for the framework's accuracy and utility. Overall, this approach offers a standardized, data-driven solution for translating multivariate diagnostic information into clinically actionable categories.

Keywords Feature Engineering | Statistical Tests | Copula Function | Machine Learning | Bootstrapping | Clustering

[†]These authors contributed equally to this work.

1 Introduction

A fundamental challenge in medical diagnostics lies in translating continuous biological variation into discrete clinical categories while preserving crucial information about underlying functional relationships. This paper addresses this challenge by developing a novel statistical framework that transforms the diagnostic assignment problem itself into a statistical decision process. As such, this work focuses on developing a methodology that accurately assigns multivariate diagnostic test results to discrete disease severity or treatment categories that can help address this major challenge across clinical practice. This difficulty is amplified by the move toward automated diagnostics and the push for standardized, statistically verifiable decision-making. Mapping diagnostic outputs to disease categories is complex and often cannot be expressed in closed form, even by clinical experts.

In this work, we address this challenge by reframing it as a three-stage multivariate statistical decision process. Our method requires a sample of patient diagnostics, including typical test results and expert-labeled disease severity assessments, which may also indicate no disease.

First, diagnostic screenings produce a multivariate set of physiological responses for a specific condition. Second, we transform these outputs into an abstract feature space through (1) feature construction, (2) feature ranking based on hypothesis testing for discriminatory power, and (3) bootstrapping to balance sample sizes across severity categories.

Finally, we learn a partitioning map from feature space to discrete disease categories, enabling systematic, automated assignment for new patients. This framework aims to standardize and improve the reliability of mapping diagnostic data to disease severity.

A detailed real data case study will be undertaken using hearing loss assessment as a motivating example. This clinical domain affects approximately 430 million individuals requiring rehabilitation [1], with projections reaching 700 million by 2050 [2, 3].

The tension between continuous measurements and discrete diagnostic categories has been extensively studied [4, 5], particularly in medical contexts where nuanced biological variation must inform actionable clinical decisions. While traditional approaches often rely on threshold-based categorization [6, 7], such methods struggle to capture both phenotype variation [8, 9] and functional outcomes [10, 11, 12] while maintaining clinical utility.

Among medical domains, hearing assessment offers a particularly clear illustration of this tension between continuous and discrete measures. In this field, different measurement types capture complementary aspects of auditory function. Pure Tone Average (PTA) provides a basic map of hearing sensitivity by averaging thresholds at key speech frequencies (typically 500, 1000, 2000, and 4000 Hz), naturally aligning with the discrete severity categories used in clinical practice and our statistical framework. However, comprehensive auditory function encompasses not just sound detection but speech comprehension, particularly in challenging environments. This broader functionality is captured through two key speech recognition measures. The Speech Reception Threshold in quiet (SRT_Q) measures the increase in decibels (dB) above normal threshold needed for 50% speech recognition under optimal conditions (distinct from the hearing thresholds measured in audiograms). The Speech Recognition Threshold in noise (SRT_N) assesses speech understanding in background noise, providing suprathreshold information beyond audibility. SRT_N indicates the signal-to-noise ratio (SNR) required for 50% intelligibility. The relationship between these measurements reveals complex patterns that resist simple categorization [13, 14], suggesting the need for a more sophisticated statistical approach to classification [15, 16].

The combination of multiple measurement types with complex interdependencies [17, 18] makes hearing assessment an ideal context for developing generalizable statistical methodology. The challenges encountered here - integrating continuous and discrete measures while preserving clinical utility - mirror those found across medical diagnostics. By transforming these measurements into a space of statistical contrasts, we can better capture the underlying structure of hearing loss while maintaining the clinical utility of discrete categories. This approach addresses fundamental challenges in medical classification that extend beyond audiology to any domain where continuous biological variation must inform discrete clinical decisions.

Traditional approaches to medical classification have relied primarily on threshold-based categorization [19], which fundamentally limits their ability to capture complex measurement structures. While advances in cut-point determination [20, 21] and maximally selected statistics [22] have improved these methods, they remain constrained by working with raw measurements rather than transformed features.

The challenge in medical classification, particularly evident in hearing assessment, manifests in three critical dimensions. First, multiple correlated measurements must be integrated while preserving their relationships [23, 24]. Second, different measurement types - from cellular-level characteristics to functional outcomes [8] - must be reconciled within

a unified framework. Third, multiple sources of uncertainty, including natural response variation and measurement precision [25], must be properly quantified [26].

Recent methodological advances have approached these challenges from different angles. Flexible modelling approaches [27] and optimal threshold methods [28] have enhanced our understanding of clinical categories. Machine learning methods [13, 14] have shown promise in pattern recognition, though often at the cost of interpretability. State-space modelling [29] has successfully captured complex dynamics between audiological measurements, while manifold learning approaches [30] have provided insights into underlying data structure.

However, a fundamental limitation persists across these approaches: they typically operate within the original measurement space rather than transforming it to better capture discriminative information. Even recent statistical learning methods [31, 32], which examine relationships between measurements under varying conditions [33, 12], focus on finding patterns rather than fundamentally reconceptualizing the feature space itself.

This limitation suggests the need for a novel approach: transforming the classification problem into a feature space where each dimension represents a statistical contrast between categories. Such a transformation would naturally integrate multiple measurement types while maintaining clinical interpretability, bridging the gap between statistical sophistication and practical utility.

Drawing upon these methodological foundations, we propose a novel statistical framework that fundamentally reimagines audiological classification. Our key insight is to transform the classification problem itself into a feature space where each dimension represents a statistical contrast between hearing loss categories, moving beyond traditional approaches that merely combine or transform raw measurements.

The framework comprises three integrated components. First, we develop a feature engineering methodology that maps audiological measurements into a higher-dimensional space of statistical contrasts. This mapping captures both the discrete nature of audiometric thresholds and the continuous characteristics of speech recognition outcomes, while naturally accommodating their complex interdependencies [34, 35]. Second, we employ complementary clustering approaches - centroid-based and hierarchical - to identify natural groupings in this statistical feature space [36, 12]. Third, we provide comprehensive validation methods that bridge statistical rigour with clinical utility [37, 38].

Our methodology builds upon statistical learning theory [39, 40] and optimal threshold determination [28], but moves beyond traditional categorization approaches by transforming the very nature of the feature space. By combining bootstrap-based feature generation with clustering machine learning techniques, we maintain strong connections to established clinical categories while revealing more nuanced patterns in hearing loss progression. To formalize this approach, we first establish notation for describing audiological measurements and their statistical transformations.

1.1 Notation & Structure

Uppercase notation denotes random quantities such as random variables, while lowercase denotes realizations of these variables obtained from measurements. Bold face indicates vectors, and non-bold face represents scalars or matrices. Sub-scripts index dimensions of arrays or sets.

The following notation is used when describing audiological measurements and their analysis. Denote by $\mathbf{X}_{N\times D}$ the random variables for the measurements for the complete set of data and its realisations from the observed experiments $\mathbf{x}_{N\times D}$, where N represents the total number of patients considered in the study, the overall sample size, and D the number of attributes collected (13 in total: pure tone thresholds at 11 frequencies [125, 250, 500, 750, 1000, 2000, 4000, 6000, 8000 Hz], speech recognition thresholds in quiet and noise). The observed attributes will be mapped into d features, where $d \geq D$, obtained from the transformations of these D observed attributes. The experimental trial data sets are taken from $G = 5$ different labelled groups of participants, with the g-th groups data, comprised of n_g participants each having recorded D observations from the audiological test battery, denoted by $\{\mathbf{x}^{(g)}_{n_g\times D}\}$, where the j-th participant in group g has observation vector $\mathbf{x}^{(g)}_j = \left[x^{(g)}_{j,1}, x^{(g)}_{j,1}, \ldots, x^{(g)}_{j,D}\right]$. Note that $n_1+n_2+n_3+n_4+n_5 = N$. Further, denote $\mathcal{G} = \{1, \ldots, G\}$ the set of groups such that $g \in \mathcal{G}$ for every g. The five groups corresponded to: group 1 ($g = 1$) participants classified with slight hearing loss (16-25 dB HL); group 2 ($g = 2$) participants classified with mild hearing loss (26-40 dB HL); group 3 ($g = 3$) participants classified with moderate hearing loss (41-60 dB HL); group 4 ($g = 4$) participants classified with moderately severe hearing loss (61-80 dB HL); and group 5 ($g = 5$) participants classified with severe hearing loss (>81 dB HL). The population mean and standard deviation for the g-th group attribute d are denoted by $\mu^{(g)}_d$ and $\sigma^{(g)}_d$, respectively. The sample estimators for these quantities will be denoted by $\widehat{\mu}^{(g)}_d$ and $\widehat{\sigma}^{(g)}_d$. The pooled variance estimator, will be denoted by S^2_p, and $F_{X^{(g)}_d(u)}$ denotes the cumulative density function (cdf) of the random variable $X^{(g)}_d$, referring to attribute d of the g-th group of participants.

The paper structure proceeds as follows: Section 2 outlines the main methodological contributions. Section 3 details the statistical framework, including the hypothesis tests used for feature screening, dependence and concordance measures, parametric and non-parametric bootstrapping techniques for feature engineering, and the clustering algorithms. Section 4 presents the real-world audiological dataset and describes its properties. Section 5 provides the empirical results, including feature selection patterns and clustering performance. Section 6 concludes with a discussion of the implications, limitations, and potential extensions of the work.

2 Contributions

Whilst the methodology presented can be applicable as a framework for many clinical diagnostic settings, the emphasis and illustration of the proposed framework will be made in the setting of audiological hearing testing to illustrate its effectiveness in this manuscript. As such, this study proposes a statistical framework for enhanced classification of hearing loss categories using multiple audiological tests. By applying specialized statistical machine learning methods for feature extraction and inference, we identify discriminative information critical for improving diagnostic accuracy. This addresses a key limitation of traditional assessments, which often rely solely on pure tone averages and may overlook broader aspects of auditory function. The main contributions are summarized below:

- An unsupervised computational method is developed to classify hearing loss categories based on multiple audiological measures. Unlike conventional approaches, it integrates diverse tests, remains robust to small and imbalanced samples, enables interpretable feature selection, and can be adapted to both supervised and unsupervised tasks. The method provides clinically relevant insights into markers of hearing loss.
- Strong clustering performance is demonstrated, using pure tone thresholds and speech recognition scores. Simulations with larger sample sizes confirm the high discriminatory power of the selected features.
- The most effective audiological predictors are identified, with pure tone thresholds (500–4000 Hz) and speech recognition in noise emerging as key contributors to differentiating hearing loss severity.

As shown in Figure 1, the framework involves three main steps. Although fundamentally unsupervised, evaluation with labelled data validates its effectiveness.

First, feature engineering constructs a mapping $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d'}$ $(d' > d)$, transforming raw audiological measurements into a higher-dimensional space. This is achieved via: **a** discriminative subspace identification through pairwise contrasts, **b** feature ranking to manage dimensionality, and **c** feature selection with bootstrapped realizations for clustering. The embedding $\varphi(\cdot)$ uses a suite of univariate and multivariate test statistics capturing frequency-specific thresholds, speech recognition patterns, and their interactions. To address the $n \ll d'$ challenge, bootstrap techniques [41] are employed, and both parametric and non-parametric tests are assessed [42, 43, 44], improving ranking accuracy and effective sample size.

Second, clustering algorithms are applied to the engineered feature space. K-Means and Hierarchical Clustering with Ward's Method [45, 46] are used due to their complementary strengths. K-Means provides efficient partitioning, while Ward's method captures hierarchical structure, important given class imbalance. Clustering performance is assessed using the Silhouette score [47], which quantifies cohesion and separation across hearing loss categories, offering insights into subcategory structure and clinical utility.

Interpretation relative to audiological attributes is discussed. Full reproducibility is ensured through the codebase available at `https://github.com/mcampi111/StatFeatEngi_Select_Clustering_Health_Severity`, which details: (1) feature engineering processes, (2) statistical test implementations, (3) clustering algorithms, and (4) evaluation and visualization tools.

3 Methods

This section details each component of the proposed methodology, beginning with the statistical tests that form the foundation of our feature space, followed by the feature engineering process that transforms audiological measurements into discriminative representations, and concluding with the clustering techniques that enables classification while maintaining clinical interpretability. Our approach combines established statistical techniques with novel adaptations specifically designed for audiological data structures, providing a framework that is both rigorous and practically applicable.

3.1 Statistical Tests

The selected hypothesis tests, used to screen for relevant features, and their test statistics are presented in Table 1. For each test we provide brief details regarding: the quantity tested; the name of the test; the null and alternative hypotheses (H_0 and H_1); the test statistic; the distribution of the statistics under the null; and degrees of freedom where appropriate.

The objective of this stage of testing is to identify which sample quantities are significantly different and can act as discriminatory features to distinguish between the disease state categories. In the illustration application in this work, it will correspond to hearing loss categories (slight, mild, moderate, moderately severe, and severe; these categories will be formally defined in Section 4) on the collected measurement variables. Table 1 presents the tests with respect to two general groups (for simplicity and without loss of generality) as group i and group j, corresponding to one pairwise combination between hearing loss categories (slight vs mild, slight vs moderate, etc.). These tests aim to screen or select relevant test statistics that will eventually be used in the feature space embedding that then gets studied in an unsupervised clustering method.

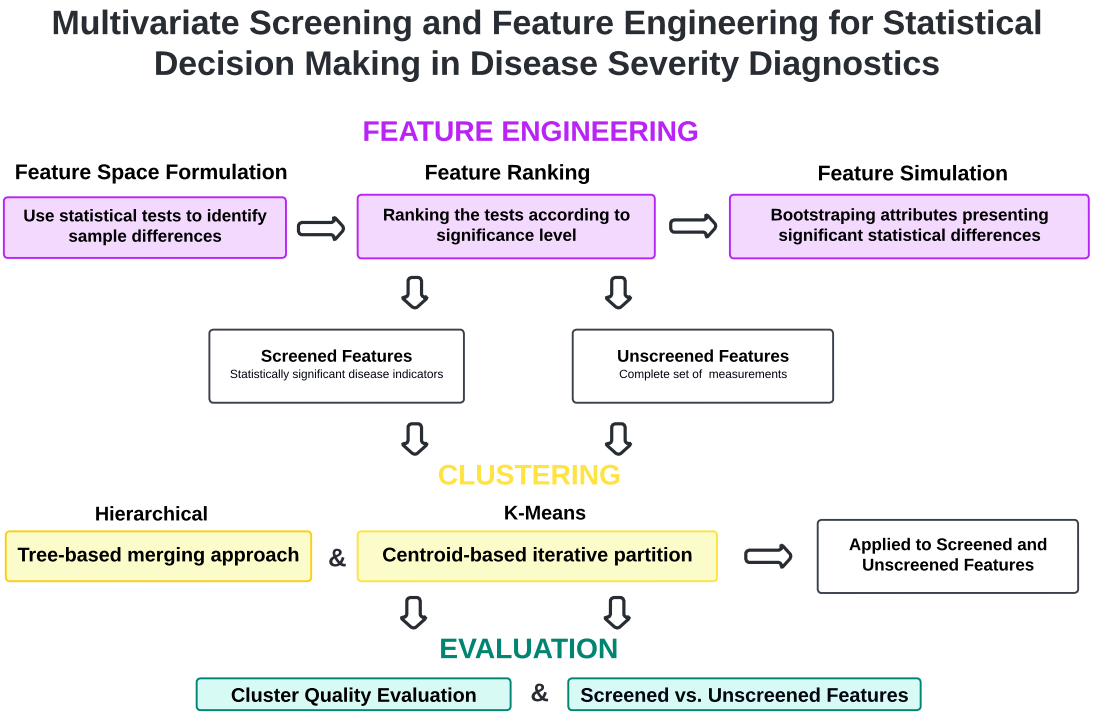

Figure 1: Overview of the proposed statistical decision framework. The method consists of two main stages for transforming raw audiological measurements into diagnostic groupings. **(1) Feature Engineering**: constructs a mapping $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d'}$ that transforms raw measurements into a higher-dimensional space via (a) discriminative subspace identification using pairwise contrasts, (b) feature ranking to reduce dimensionality, and (c) bootstrapped realizations to improve sample balance. The resulting screened features are statistically selected and transformed attributes. **(2) Clustering and Evaluation**: centroid-based and hierarchical clustering are applied to both screened and unscreened features. Evaluation includes (i) comparison of cluster quality (e.g., silhouette score), (ii) assessment of alignment with diagnostic labels, and (iii) benchmarking screened features against raw inputs. "Unscreened features" are original audiological measurements; "screened features" are selected transformations.

The first type of test considered searches for mean sample differences for pairwise combinations and variables. We employ both the standard t-Student mean diffrence test, considering equality in variances between the tested samples, and the Welch's t-test [48], which accommodates unequal variances between sample groups. The second class of tests targets sample variance using the variance ratio test (F-test) [49] and the Bartlett Test [50], which assess whether variances across different groups can be considered equal. The third class of tests examines distribution sam-

ple differences using the Kolmogorov-Smirnov test [51], which is sensitive to differences in location and shape of empirical distribution functions, and the Cramer-von-Mises test with different choices of weighting functions [52], in which for appropriate choice of quantile weighting function can perform particularly well for detecting differences in heavy-tailed distributions (see the Supplementary Appendix for detailed formulation).

Beyond univariate tests, we employ several multivariate approaches to examine interdependencies between audiological measurements. The sparse covariance matrix comparison method [53] is particularly relevant for detecting subtle differences between adjacent hearing loss categories where covariance differences may be sparse. Tukey's Honestly Significant Difference (HSD) test [54] provides controlled pairwise mean comparisons across multiple groups, while controlling family-wise error rates. For capturing complex dependence structures between different audiological measurements, we employ the Copula test [55], which evaluates equality between dependence structures while excluding marginal behaviours (detailed mathematical formulation is provided in the Supplementary Appendix).

3.1.1 Dependence & Concordance Measures

While the copula test provides insights into the overall dependence structure, additional measures of dependence and concordance can offer complementary information about the relationships between the medical diagnostic test array outputs, in the application setting considered these will be audiological measurements. They capture different aspects of the dependence structure between different measurement coordinates (tests) and remain invariant under monotone transformations of the data [55, 56, 57].

For any pair of measurements $(X_{ld}, Y_{md'})$, where l, m index observations and d, d' index attributes (which could be the same or different frequencies, or speech recognition scores) from groups i and j respectively, with corresponding sample sizes n_i and n_j, we transform the data to ranks:

$$U_{ld,n_i} = \frac{\text{rank}(X_{ld})}{n_i + 1}, \quad V_{md',n_j} = \frac{\text{rank}(Y_{md'})}{n_j + 1}.$$

We then compute several copula-based measures, which are summarized in Table 2. In each formula, the expectation $\mathbb{E}[\cdot]$ is taken over the empirical distribution of the rank-transformed variables $(U_{ld,n_i}, V_{md',n_j})$.

These measures collectively provide a comprehensive view of dependence structures in audiological data. The modified Kendall's tau and Spearman's rho offer robust assessments of general relationships, while the multivariate and sign-based measures capture more nuanced patterns. The absolute difference and Gini-based measures are particularly valuable for identifying discrepancies and extreme patterns, and the local Gaussian correlation provides insight into the dependence structure via a Gaussian transform. Finally, the tail-dependence measures λ_L and λ_U highlight lower- and upper-tail co-movements that can be especially relevant for extremes in hearing measurements.

To understand the complexity of these dependency measures, we provided a visualization of different types of statistical relationships in the Supplementary Appendix (see Figure 1). In audiological datasets, hearing loss progression and speech recognition abilities often exhibit intricate, non-linear relationships that traditional statistical methods might fail to detect unless copula based methods are adopted beyond the simple Guassian copula setting. For instance, pure-tone thresholds at different frequencies, speech recognition scores in quiet and noisy environments, and age-related hearing changes can interact in complex ways—not always following simple linear correlation patterns. The figure illustrates various dependency structures, including strong and moderate linear relationships, upper and lower tail dependencies, nonlinear patterns, asymmetric interactions, and threshold effects. These diverse patterns are particularly relevant in hearing assessment, where subtle interactions between different audiometric measurements can provide crucial diagnostic insights. By using copula-based measures like Modified Kendall's tau, Spearman's rho, and tail dependence indicators, we can capture nuanced relationships that might indicate progressive hearing loss, individual variability in hearing function, or unique audiological profiles that would be invisible to traditional statistical approaches.

Using the statistical tests above described, we performed comprehensive pairwise comparisons across all hearing loss categories (slight vs mild, slight vs moderate, etc.) for each audiological measurement. Additionally, we conducted aggregate comparisons across multiple categories simultaneously to capture broader patterns of discriminative power.

3.2 Feature Engineering

The proposed feature engineering approach builds upon the results of statistical significance testing, using both parametric and non-parametric bootstrapping techniques [58] to generate bootstrapped features for the attributes that show discriminative power between hearing loss categories. These statistically significant differences serve as ideal candidates for our feature space formulation precisely because they capture the most informative contrasts between severity categories, providing the foundation for our statistical decision framework. When statistical tests identify significant differences between groups for particular attributes, we apply a systematic bootstrapping procedure to generate

Statistical Tests

Univariate Tests

Feature	Test	H_0	H_1	Test Statistic	Distribution & DOF
Mean	T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{S_p^2 \sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Student's t, DOF: $n_g + n_h - 2$
Mean	Welch T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{\sqrt{\frac{S_d^{2(g)}}{n_g} + \frac{S_d^{2(h)}}{n_h}}}$	Student's t, DOF: Welch–Satterthwaite
Variance	Variance Ratio	$\sigma_d^{2(g)} = \sigma_d^{2(h)}$	$\sigma_d^{2(g)} \neq \sigma_d^{2(h)}$	$F = \frac{S_d^{2(g)}}{S_d^{2(h)}}$	Fisher–Snedecor DOF: $F_{(n_g-1,\ n_h-1)}$
Distr.	Kolmogorov Smirnov	$F_d^{(g)}(x) = F_d^{(h)}(x)$	$F_d^{(g)}(x) \neq F_d^{(h)}(x)$	$D = \sup_x \left\lvert \hat{F}_d^{(g)}(x) - \hat{F}_d^{(h)}(x) \right\rvert$	Free (no DOF)
Distr.	Cramer-Von Mises	$F_d^{(g)}(x) = F_d^{(h)}(x)$	$F_d^{(g)}(x) \neq F_d^{(h)}(x)$	$Q = n_g n_h \int_{-\infty}^{\infty} w(x) \left[\hat{F}_d^{(g)}(x) - \hat{F}_d^{(h)}(x)\right]^2 d\hat{F}_d^{(h)}(x)$	Free (no DOF)

Multivariate Tests

Feature	Test	H_0	H_1	Test Statistic	Distribution & DOF
Variance	Bartlett Test	$\sigma_d^{2(g)} = \sigma_d^{2(h)}$ $\forall\ (g,h)$	$\sigma_d^{2(g)} \neq \sigma_d^{2(h)}$ for at least one (g,h)	$T = \frac{(N-G)\ln(S_p^2) - \sum_{l=1}^{G}(n_g-1)\ln(S_d^{2(g)})}{1 + \left(\frac{1}{3(G-1)}\right)\left[\left(\sum_{l=1}^{G}\frac{1}{n_g-1}\right) - \frac{1}{N-G}\right]}$	Chi-Square $\chi^2_{(G-1)}$
Covariance		$\mathbf{\Sigma}_g = \mathbf{\Sigma}_h$	$\mathbf{\Sigma}_g \neq \mathbf{\Sigma}_h$	$M_n - 4\log p + \log\log p$	Type I extreme value (no DOF)
Tukey	Tukey HSD	$\mu_d^{(g)} = \mu_d^{(h)}$ $\forall\ (g,h)$	$\mu_d^{(g)} \neq \mu_d^{(h)}$ for at least one g, h	$W = \frac{\max_{(g,h)}\left(\bar{X}_d^{(g)} - \bar{X}_d^{(g)}\right)}{\sqrt{\frac{1}{2}\frac{S_d^{2(g)} + S_d^{2(h)}}{n_{g,d}+n_{h,d}}}}$	Studentized range DOF: $q_{(G,\ N-G)}$
Copula		$C_g = C_h$	$C_g \neq C_h$	$E_{n_g,n_h} = \frac{\hat{C}_g - \hat{C}_h}{\sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Free (no DOF)

Table 1: Descriptions of the considered statistical tests for feature screening and selection (stage 1). The sample estimators required for the implementation of the tests are given as follows (following the order of the introduced tests). $\bar{X}_d^{(g)}, \bar{X}_d^{(h)}$ represents the mean estimators. S_p^2 corresponds to the pooled variance estimator. $S_d^{2(g)}, S_d^{2(h)}$ are the variance estimators and n_g, n_h the correspondent groups sample sizes. N is the total sample size. S_d^2 is variance estimator for group g, n_g is the sample size of the g-th group, G represents the number of groups and S_p^2 the pooled variance estimator corresponding to $S_p^2 = \sum_{l=1}^{k}(n_l - 1)(S_l^2)/(n-k)$. $F_d^{(g)}(x), F_d^{(h)}(x)$ are the cumulative distribution functions of the two groups and $\hat{F}_d^{(g)}(x), \hat{F}_d^{(h)}(x)$ are the empirical cumulative distribution functions, respectively. $\bar{X}_d^{(h)}$ is larger of the means and $\bar{X}_d^{(g)}$ is smaller of the means. In the case of the Tukey test, HSD stands for Honest Significant Difference.

Measure	Formula	Description
Modified Kendall's tau	$\tau_{\text{cop}} = 4\,\mathbb{E}[U_{ld,n_i}\, V_{md',n_j}] - 1$	Captures overall concordance patterns; sensitive to monotonic relationships across frequencies.
Modified Spearman's rho	$\rho_{\text{cop}} = 12\,\mathbb{E}\Big[\, U_{ld,n_i}\, \dfrac{l-0.5}{n_i}\Big] - 3$	Rank-based correlation measure for gradual changes in hearing function.
Multivariate Spearman's rho	$\rho_{\text{multi}} = 12\,\mathbb{E}[(U_{ld,n_i} - 0.5)\,(V_{md',n_j} - 0.5)]$	Captures complex dependencies between different aspects of hearing function.
Sign-based association	$\beta_{\text{cop}} = 4\,\mathbb{E}\Big[\text{sign}\big((U_{ld,n_i} - 0.5)\,(V_{md',n_j} - 0.5)\big)\Big]$	Robust measure for directional relationships, less sensitive to outliers.
Concordance measure	$\gamma_{\text{cop}} = \mathbb{E}[\lvert U_{ld,n_i} - V_{md',n_j}\rvert]$	Quantifies disagreement between measurements across frequencies/tests.
Gini-based measure	$\text{Gini}_{\text{alt}} = \mathbb{E}[\lvert 2\,U_{ld,n_i} - 1\rvert\ \lvert 2\,V_{md',n_j} - 1\rvert]$	Sensitive to differences in distribution tails; identifies extreme patterns.
Local Gaussian correlation	$\rho_{\text{local}} = \text{cor}\big(\Phi^{-1}(U_{ld,n_i}),\ \Phi^{-1}(V_{md',n_j})\big)$	Provides insights into dependence structure through a Gaussian transformation.
Tail Dependence	$\lambda_L \approx \dfrac{1}{k} \sum_{q \in \mathcal{Q}_L} \dfrac{\Pr\big(U \le \widehat{F}_U^{-1}(q),\, V \le \widehat{F}_V^{-1}(q)\big)}{q},$	Approximates lower (λ_L) and upper (λ_U) tail dependence by averaging empirical joint tail probabilities. Here, $k = \lvert\mathcal{Q}_L\rvert$ or $\lvert\mathcal{Q}_U\rvert$, the number of quantiles in each tail grid.
	$\lambda_U \approx \dfrac{1}{k} \sum_{q \in \mathcal{Q}_U} \dfrac{\Pr\big(U \ge \widehat{F}_U^{-1}(q),\, V \ge \widehat{F}_V^{-1}(q)\big)}{1-q}$	Values > 1 indicate stronger co-occurrence in the respective tail than under independence.

Table 2: Summary of Copula-based Dependence Measures

balanced feature sets that capture these differences while maintaining equal representation across groups. For each significant attribute identified through statistical testing, we employ two bootstrapping methods: one parametric approach using a Normal distribution and a second non-parametric approach. This combination allows us to generate robust feature sets while accounting for different potential underlying distributions in the audiological measurements. In this subsection we present the different bootstrapping techniques used for the feature engineering stage of the proposed framework.

3.2.1 Parametric and Non-Parametric Copula and Univariate Bootstrapping

An overview of the bootstrapping methodology is present for four fundamental approaches - parametric bootstrap, non-parametric bootstrap, parametric copula bootstrap, and non-parametric copula bootstrap - with detailed insights and theoretical foundations for copula estimation and its bootstrapping procedure provided in detail in the Supplementary Appendix.

Bootstrap methods, introduced by [58], provide a simulation-based framework for assessing statistical accuracy. While multiple variations exist in both parametric and non-parametric forms, the core principle remains consistent: using resampling to estimate statistical properties when analytical solutions are impractical or collection of further samples is infeasible, invasive, too costly or impractical. In many medical diagnostic settings, the data collected may be onerous on the patient and expensive to collect. Furthermore, the proportion of the population with different diseases states may be naturally imbalanced in many medical assessment settings. The combination of these factors often results in sample sets that are used for medical diagnostic assessment that are imbalanced. This is where bootstrap methods can be directly beneficial, and especially those that consider carefully the possible multivariate concordance structures present in the data.

The fundamental bootstrap problem considers an i.i.d. sample $X_1, \ldots, X_n \sim f(x|\theta)$ and aims to estimate the standard error of an estimator $\hat{\theta}$ of θ. Traditional approaches rely on asymptotic theory to study the sampling distribution and variance of $\hat{\theta}$ for large samples. Bootstrap provides an alternative through simulation: repeatedly generating samples $X_1^*, \ldots, X_n^* \sim f(x|\theta)$, computing $\hat{\theta^*}$ for each sample, and taking the empirical standard deviation of these estimates. However, since the true parameter θ is unknown, making direct simulation impossible, bootstrap instead generates

samples from an estimate of the underlying distribution. For a statistic of interest $T := T(X_1, \ldots, X_n)$, two main approaches exist: a parametric approach that fits and samples from a specified probability distribution, and a non-parametric approach that uses the empirical distribution of the observed data.

Parametric Bootstrap: In the parametric approach, we fit a model $f(x|\theta)$ to the observed data. This is called the parametric bootstrap because the simulated data comes from a fitted parametric model $f(x|\theta)$.

Algorithm 1: Parametric Bootstrap

Input: Parametric model $f(x|\theta)$ fit to $X_1, \ldots, X_n$ using an estimator $\hat{\theta}^*$ (for example, the MLE)

for $i = 1, 2, \ldots, B$ **do**

1. Simulate iid samples $X_1^*, \ldots, X_n^* \sim f(x|\hat{\theta})$
2. Compute the statistic $T^* := T(X_1^*, \ldots, X_n^*)$ on the data $X_1^*, \ldots, X_n^*$

Output: the empirical standard deviation of T^* across the B simulations

Non-Parametric Bootstrap: The non-parametric approach takes a different perspective: instead of assuming a parametric form, it estimates the true distribution using the empirical distribution of the data, which places mass $\frac{1}{n}$ at each observed value $X_1, \ldots, X_n$. Under this approach, generating i.i.d. samples $X_1^*, \ldots, X_n^*$ amounts to sampling with replacement from the original data. Note that this typically results in repeated values in the resampled data, even when the original observations were all distinct. The procedure is as follows: This approach requires no assumptions on

Algorithm 2: Non-Parametric Bootstrap

Input: $X_1, \ldots, X_n$

for $i = 1, 2, \ldots, B$ **do**

1. Simulate iid samples $X_1^*, \ldots, X_n^*$ as n samples with replacement from original data $X_1, \ldots, X_n$
2. Compute the statistic $T^* := T(X_1^*, \ldots, X_n^*)$ on the data $X_1^*, \ldots, X_n^*$

Output: the empirical standard deviation of T^* across the B simulations

a particular form of parametric model. An alternative approach to traditional bootstrapping methods is copula-based bootstrapping, which allows for preserving dependence structures in multivariate data. A copula function captures the joint dependence between variables while allowing marginal distributions to be modelled separately. This method is particularly useful for high-dimensional data where traditional bootstrapping may not effectively preserve dependencies.

Parametric Copula Bootstrap The parametric version of copula-based bootstrapping assumes that the joint distribution of $(X_1, \ldots, X_d)$ can be factored through a parametric copula. In particular, we assume

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \;=\; C_\theta\big(F_1(x_1), \ldots, F_d(x_d)\big),$$

where $F_j(\cdot)$ is the marginal distribution function of X_j, and C_θ is a t-copula with parameters $\theta = (R, \nu)$, comprising the correlation matrix R and degrees of freedom ν. Let $X_{n\times d}$ be our data matrix of n observations and d variables.

Non-Parametric Copula Bootstrap: The non-parametric version does not assume a functional form for the copula. Instead, one estimates a flexible Bernstein copula $\widehat{C}(\cdot)$ from the rank-transformed data $\big(\widehat{F}_1(x_{i1}), \ldots, \widehat{F}_d(x_{id})\big) \in [0,1]^d$. Let $X_{n\times d}$ again be our data matrix.

The copula-based bootstrap methods effectively capture dependencies in multivariate data while allowing flexibility in marginal distributions. The parametric approach benefits from the robustness of the t-copula in capturing tail dependencies, while the non-parametric Bernstein copula offers greater flexibility without parametric assumptions. Note that the sampling procedure for this copula is described in the Supplementary Appendix.

3.2.2 Screening-Specific Bootstrap for Feature Construction

The combination of the statistical tests combined with the application of the parametric and non-parametric bootstrapping methods applied to the relevant subsets of the data being tested allows for the accurate selection and screening of

Algorithm 3: Parametric Student-t Copula Bootstrap

Input: Multivariate data matrix $X_{n\times d} = \{(x_{i1}, \ldots, x_{id})\}_{i=1}^{n}$; Fitted t-copula parameters $\theta = (R, \nu)$; Marginal distribution estimates $\hat{F}_1, \ldots, \hat{F}_d$ (e.g., via MLE or empirical CDFs).

for $i = 1, 2, \ldots, B$ **do**

1. **Sample from the t-copula:** Generate n independent samples
$$(u_{i1}, \ldots, u_{id}) \sim tCopula(R, \nu), \quad i = 1, \ldots, n.$$
Each $(u_{i1}, \ldots, u_{id})$ lies in $[0, 1]^d$.
2. **Structure the samples:** Organize the samples into vector form
$$U_1^*, \ldots, U_n^*, \text{ where each } U_i^* = (u_{i1}, \ldots, u_{id})$$
3. **Compute statistic of interest:** Evaluate
$$T^* := T\big(U_1^*, \ldots, U_n^*\big).$$

Output: empirical distribution (or empirical moments) of T^* based on the B replicates.

Algorithm 4: Non-Parametric Bernstein Copula Bootstrap

Input: Multivariate data matrix $X_{n\times d} = \{(x_{i1}, \ldots, x_{id})\}_{i=1}^{n}$; A non-parametric Bernstein copula fit $\widehat{C}$, Marginal distribution estimates $\hat{F}_1, \ldots, \hat{F}_d$ (often empirical).

for $i = 1, 2, \ldots, B$ **do**

1. **Estimate copula from rank-transformed data:** Using $(\hat{F}_j(x_{ij}))_{i=1,\ldots,n,\, j=1,\ldots,d}$, fit a Bernstein copula $\widehat{C}$ to approximate scale free dependence .
2. **Sample from the estimated copula:** Generate n samples
$$(u_{i1}, \ldots, u_{id}) \sim \widehat{C}, \quad i = 1, \ldots, n.$$
3. **Structure the samples:** Organize the samples into vector form
$$U_1^*, \ldots, U_n^*, \text{ where each } U_i^* = (u_{i1}, \ldots, u_{id})$$
4. **Compute statistic of interest:** Evaluate
$$T^* := T\big(U_1^*, \ldots, U_n^*\big).$$

Output: empirical distribution (or empirical moments) of T^* based on the B replicates.

discriminatory features from the diagnostic test data features. As such, the method aims to generate n' new samples for each group g to form an augmented feature vector that will be used in the assessment and screening.

Bootstrap Setup Remark that $X_{g,(d_k)}$ is the observed data for group $g \in \mathcal{G}$ and attribute d_k. Tnen $N' = 5 \times n'$ new samples (since we resample for *all* 5 groups, ensuring balanced representation) under three different bootstrap approaches is utilised to both determine whether the contrast between some groups i and j feature(s) are statistically significant for d_k. The procedure is outlined as follows:

1. Parametric (Normal Distribution): Model $X_{g,(d_k)}$ as $\mathcal{N}\big(\hat{\mu}_g, \hat{\sigma}_g^2\big)$, where $(\hat{\mu}_g, \hat{\sigma}_g^2)$ are estimated from the observed data (e.g., method of moments).
2. Non-Parametric: Sample n' times *with replacement* directly from $X_{g,(d_k)}$, treating its empirical distribution as the underlying model.

In each case, denote the newly generated samples for group g as $\{x_{1,(d_k)}^{*(g)}, x_{2,(d_k)}^{*(g)}, \ldots, x_{n',(d_k)}^{*(g)}\}$.

Feature Construction for Univariate Tests For each type of hypothesis test for which a statistically significant structure was detected, we then derive a corresponding set of summary statistics as new features from the screened data structure given as follows:

- **Mean Test (e.g., t-test)**: Compute bootstrapped means $\{\bar{x}^{(g)}_{1,(d_k)}, \ldots, \bar{x}^{(g)}_{n',(d_k)}\}$ for each group $g \in \mathcal{G}$. Concatenate across groups to form

$$\widetilde{\mathbf{x}}_{N' \times 1} \;=\; \big[\bar{x}^{(1)}_{1,(d_k)}, \ldots, \bar{x}^{(1)}_{n',(d_k)}, \;\ldots,\; \bar{x}^{(5)}_{1,(d_k)}, \ldots, \bar{x}^{(5)}_{n',(d_k)}\big]^\top.$$

- **Variance Test**: Compute bootstrapped variances for each group; similarly concatenate them into an $N' \times 1$ vector.

$$\widetilde{\mathbf{x}}_{N' \times 1} \;=\; \big[s^{2^{(1)}}_{1,(d_k)}, \ldots, s^{2^{(1)}}_{n',(d_k)}, \;\ldots,\; s^{2^{(5)}}_{1,(d_k)}, \ldots, s^{2^{(5)}}_{n',(d_k)}\big]^\top$$

- **Distribution Test (e.g., KS or CvM)**: Compute robust summary statistics, such as median, interquartile range (IQR), and the 5th/95th percentiles, from each bootstrapped sample. Concatenate these into feature vectors per group.

$$\widetilde{\mathbf{x}}_{N' \times 4} \;=\; \left[\left(\text{med}^{(1)}_{1,(d_k)} \quad \text{IQR}^{(1)}_{1,(d_k)} \quad p5^{(1)}_{1,(d_k)} \quad p95^{(1)}_{1,(d_k)}\right), \ldots, \left(\text{med}^{(5)}_{n',(d_k)} \quad \text{IQR}^{(5)}_{n',(d_k)} \quad p5^{(5)}_{n',(d_k)} \quad p95^{(5)}_{n',(d_k)}\right)\right]^\top$$

Repeating this procedure for each significant attribute d_k produces a new feature d'_k. When multiple attributes are significant, their corresponding bootstrapped statistics yield a set of new attributes $\{d'_1, d'_2, \ldots\}$. This process is repeated for each bootstrap method (Normal, Non-Parametric) to form three separate feature matrices.

Feature Construction for Multivariate Tests - Copula For multivariate relationships—identified, for example, via copula-based dependence tests between d_k of group g and d_l of group h—we generate two additional categories of features.

- **Rank Transformations:** For each newly bootstrapped sample, transform $(x^{*(g)}_{(d_k)}, x^{*(h)}_{(d_l)})$ into rank scale $U^{(g)}_{(d_k)}, U^{(h)}_{(d_l)}$, yielding vectors $\widetilde{\mathbf{x}}^{\text{rank}}_{N' \times 2}$ when concatenating across n' samples and groups.

$$\widetilde{\mathbf{x}}^{\text{rank}}_{N' \times 2} \;=\; \big[\left(U^{(g)}_{1,(d_k)} \quad U^{(h)}_{1,(d_l)}\right), \ldots, \left(U^{(g)}_{n',(d_k)} \quad U^{(h)}_{n',(d_l)}\right)\big]^\top$$

- **Copula-Based Dependence Measures:** Depending on the bootstrap approach:
  - Parametric Case: We use a t-copula for generating multivariate samples, preserving heavy-tailed dependence.
  - Non-Parametric Case: We use a Bernstein copula fitted to the observed rank-transformed data.

  From the bootstrapped samples of (d_k, d_l), compute the dependence measures corresponding to Table 2:

$$\tau_{\text{cop}},\; \rho_{\text{cop}},\; \rho_{\text{multi}},\; \beta_{\text{cop}},\; \gamma_{\text{cop}},\; \text{Gini}_{\text{alt}},\; \rho_{\text{local}},\; \boldsymbol{\lambda} = (\lambda_L, \lambda_U).$$

  These measures form feature vectors of dimension $N' \times M$, where M is the total number of dependence statistics. We denote the resulting feature matrix as:

$$\widetilde{\mathbf{x}}^{\text{cop}}_{N' \times M} \;=\; \left[\left(\tau^{(g,h)}_{\text{cop},1} \quad \rho^{(g,h)}_{\text{cop},1} \quad \cdots \quad \lambda^{(g,h)}_{L,1} \quad \lambda^{(g,h)}_{U,1}\right), \ldots, \left(\tau^{(g,h)}_{\text{cop},n'} \quad \rho^{(g,h)}_{\text{cop},n'} \quad \cdots \quad \lambda^{(g,h)}_{L,n'} \quad \lambda^{(g,h)}_{U,n'}\right)\right]^\top$$

In this manner, the chosen copula model (i.e., t-copula or Bernstein copula) corresponds to the parametric or non-parametric approach, respectively, ensuring a consistent methodology for high-dimensional dependence modelling.

Feature Construction for Multivariate Tests - Covariance If a test indicates that $\text{Cov}(d_k, d_l)$ is significant for groups g and h, we similarly bootstrap n' samples for d_k (group g) and d_l (group h) while also sampling from all other groups to maintain consistency. If variances for either attribute have not already been bootstrapped (i.e., they were not flagged in univariate variance tests), we compute n' correlations $\{\rho^{*(g)}_{(d_k,d_l)}, \rho^{*(h)}_{(d_k,d_l)}\}$ for each group. These are concatenated into a feature matrix

$$\widetilde{\mathbf{x}}_{N' \times 2} \;=\; \left[\left(\rho^{(g)}_{1,(d_k,d_l)} \quad \rho^{(h)}_{1,(d_k,d_l)}\right), \ldots, \left(\rho^{(g)}_{n',(d_k,d_l)} \quad \rho^{(h)}_{n',(d_k,d_l)}\right)\right]^\top$$

Feature Construction for Multivariate Tests - Bartlett For the Bartlett test comparing variances across all groups simultaneously, we compute bootstrapped variances and their ratios. For each attribute d_k found significant by the Bartlett test, we compute variances for all groups and form:

$$\widetilde{\mathbf{x}}_{N' \times G} \;=\; \left[\left(s^{2^{(1)}}_{1,(d_k)} \quad s^{2^{(2)}}_{1,(d_k)} \quad \cdots \quad s^{2^{(G)}}_{1,(d_k)}\right), \ldots, \left(s^{2^{(1)}}_{n',(d_k)} \quad s^{2^{(2)}}_{n',(d_k)} \quad \cdots \quad s^{2^{(G)}}_{n',(d_k)}\right)\right]^\top$$

Sample Sizes and Final Feature Matrices The above procedures are repeated for varying sample sizes $n' \in \{50, 500, 1000, 5000\}$ to assess the effect of bootstrap sample size. Ultimately, we obtain two feature matrices:

$$\widetilde{\mathbf{X}}^{\mathcal{N}}_{N' \times D'}, \quad \widetilde{\mathbf{X}}^{\mathrm{NP}}_{N' \times D'},$$

corresponding to Normal parametric ($\mathcal{N}$) and Non-Parametric (NP) bootstrapping, respectively. In subsequent clustering analysis, we compare performance using either screened features (derived from statistically significant variables) or unscreened features (from non-significant variables) to gauge discriminative power under each approach.

3.3 Clustering Analysis

To identify natural groupings in our engineered features that may correspond to hearing loss categories, we employ two complementary clustering approaches. First, K-means clustering provides efficient partitioning based on centroid distances, making it suitable for identifying distinct groups in our high-dimensional feature space. Second, hierarchical clustering with Ward's method captures nested relationships between groups, particularly valuable given the progressive nature of hearing loss severity and potential subgroups within traditional categories.

The K-means algorithm partitions the observations into $k = 5$ clusters (corresponding to the known hearing loss categories) by iteratively minimizing the within-cluster sum of squares. For robustness, we perform multiple restarts with different random initializations and select the solution with minimal total within-cluster sum of squares.

The hierarchical clustering with Ward's method complements this approach by building a dendrogram that reveals the nested structure of clusters. Ward's criterion minimizes the total within-cluster variance while merging clusters, making it particularly suitable for detecting subtle gradations in hearing loss severity.

Full technical details of these clustering methods are provided in the Supplementary Appendix, with comprehensive results presented in Section 5.

4 Data Description and Properties

This section is dedicated to the case study real data description. We begin by outlining the configuration of the tests administered to the participants, including the audiogram and two speech tests conducted in quiet and noise, respectively. Following this, we provide a detailed dataset description, incorporating descriptive statistics and visual representations. Specifically, we present several plots illustrating the distribution of the critical variables as violin plots that offer a deeper insight into the variability and central tendencies within the data.

4.1 Data Acquisition & Testing Procedures

Our study utilizes a dataset from Amplifon France, which contains routine data from hearing aid fitting practices across multiple Amplifon hearing aid acoustician labs in France. For retrospective data analysis, the dataset was provided in pseudonymized form to Institut Pasteur under the BIG DATA AP project. The data protection authority Commission Nationale de l'Informatique et des Libertés (the National Commission on Informatics and Liberty) authorised the processing of BIG DATA AP study data on April 05, 2024.

The dataset included participants' age, sex assigned at birth, pure-tone audiograms for both ears, and speech recognition thresholds for speech tests in quiet and noise, respectively. The degree of hearing loss was derived by calculating the pure-tone average (PTA) based on individual hearing thresholds at 0.5, 1, 2, and 4 kHz [59], according to the American Speech-Language-Hearing Association (ASHA) classification, detailed in Table 3.

The study focused on participants aged 40-90 years with symmetric hearing loss, defined by a PTA difference of less than 15 dB between ears [60]. This age range was selected based on data availability and completeness. The final dataset comprised 48,144 participants. Data on race or ethnicity were not collected, adhering to French legal restrictions on personal data collection per the 1978 Law on Information and Freedoms [61]. Specific details about the data acquisition procedures are provided in the Supplementary Appendix.

Pure-tone average (PTA) Categories	
Degree of hearing loss	PTA range (dB HL)
Normal	–10 to 15
Slight	16 to 25
Mild	26 to 40
Moderate	41 to 55
Moderately severe	56 to 70
Severe	71 to 90

Table 3: Pure-tone average (PTA) categories were defined in accordance with the American Speech-Language-Hearing Association (ASHA) classification [59]. Note that we do not have any participant with normal hearing in our dataset.

4.2 Data Description

This subsection describes the dataset, with Table 4 providing a comprehensive summary of descriptive statistics, including age, audiogram frequencies (125 Hz to 8000 Hz), SRT_N, and SRT_Q. The mean participant age is approximately 73 years, with a standard deviation (SD) of 9.73 years, indicating an elderly population with some variability, ranging from 40 to 90 years.

Mean hearing thresholds increase over frequency, from about 30 dB HL at 125 Hz to about 72 dB HL at 8 kHz, with standard deviations ranging from 13 to 19 dB HL. Median values are slightly below the means, suggesting a slight right skew. The database contains hearing thresholds spanning the entire range of measurable audiometric levels. The mean SRT_N is 4.43 dB SNR with a standard deviation (SD) of 3.96 dB SNR, while the mean SRT_Q is 45.97 dB SPL with an SD of 11.56 dB SPL. The SRT_Q median is close to the mean, indicating a symmetrical distribution, while the SRT_N median is slightly lower than the mean, suggesting a minor positive skew. SRT_N values range from -10 dB SNR to 20 dB SNR, and SRT_Q values range from 5 dB SPL to 80 dB SPL.

This analysis includes the entire sample, covering all degrees of hearing loss, ages 40 to 90, and both sexes.

Figure 2 illustrates the distribution of individuals across different degrees of hearing loss. The categories include slight, mild, moderate, moderately severe, and severe hearing loss. The majority of individuals fall into the Moderate and Mild categories, with nearly equal number of patients (Moderate: 20,246; Mild: 18,979). The moderately severe and slight categories have fewer individuals, with 4,826 and 3,704 people respectively. Finally, the severe category has the smallest number of individuals, with only about 389 people, showing that severe hearing loss is less common within this population. This distribution highlights the varying prevalence of hearing loss severity among the individuals studied.

Descriptive Statistics Overall

Statistics	**Age**	**Frequencies (Hz)**											SRT_N	SRT_Q
		125	**250**	**500**	**750**	**1000**	**1500**	**2000**	**3000**	**4000**	**6000**	**8000**		
Mean	72.98	30.48	31.24	33.83	36.40	38.11	45.08	48.48	55.79	61.74	70.35	71.71	4.43	45.97
Median	74.00	30.00	30.00	30.00	35.00	35.00	45.00	50.00	55.00	60.00	70.00	70.00	4.00	45.00
SD	9.73	13.19	14.50	15.14	15.43	15.68	16.18	16.28	16.50	16.97	18.24	18.75	3.96	11.56
Min	40	-10.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	-5.00	3.00	-10.00	5.00
Max	90	120.00	120.00	120.00	120.00	120.00	120.00	120.00	125.00	125.00	125.00	130.00	20.00	80.00

Table 4: Descriptive statistics over the whole sample population. Variables include age, audiogram frequencies, SRT_N and SRT_Q. Note that the unit of measures are dB HL for the audiogram frequencies, dB SNR for SRT_N and dB SPL for SRT_Q.

Figure 3 shows data distributions via violin plots. The left panel displays hearing thresholds for the left ear. The right panel illustrates SRT_N and SRT_Q values for the left ear, as these measures were similar for both ears. For brevity, descriptive statistics are reported solely for the left ear, given symmetric hearing loss was considered in our data.

In summary, hearing thresholds vary across age groups, with lower thresholds observed from 40-45 to 65-70 years, particularly at lower frequencies (125 Hz to 1000 Hz). However, thresholds increase markedly at higher frequencies with age, rising from 55.44 dB in the 40-45 age group to 81.85 dB in the 85-90 group, indicating age-related hearing loss. SRT_N and SRT_Q values also increase with age, reflecting a decline in speech perception in noisy environments: mean SRT_N rises from 2.79 dB (40-45 years) to 6.94 dB (85-90 years), while SRT_Q increases from 40.70 dB to 53.73 dB.

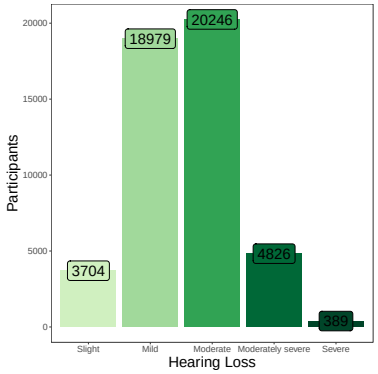

Figure 2: Sample size across Pure Tone Average (PTA) hearing loss categories for the whole database population. Hearing loss degree is classified based on PTA categories.

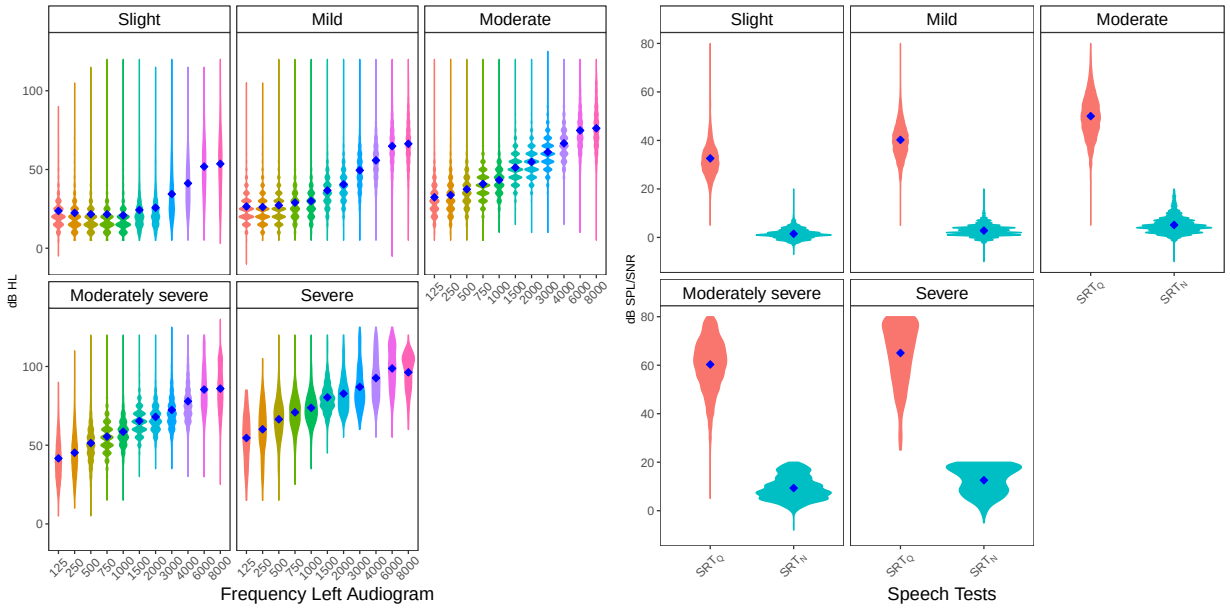

Figure 3: Violin plots of hearing thresholds at different frequencies (left) and SRT_Q, SRT_N (right) by hearing loss degree for the left ear. The x-axis on the left plot represents the measured frequencies from 125 Hz to 8000 Hz, while the x-axis on the right represents the speech tests. The y-axis shows hearing thresholds in dB HL (left) and SRT_Q and SRT_N in dB SPL/SNR (right). Due to the symmetrical nature of hearing loss in the sample, the left ear was selected for this representation. Equivalent results were confirmed when analysing the right ear, ensuring the reliability of the observed patterns.

5 Results

The results section presents a comprehensive investigation into the statistical characteristics and classification of hearing loss across different severity categories. Employing a multifaceted analytical approach, this study systematically deconstructs the complex landscape of audiological measurements. The analysis progresses through several key stages: initial exploratory data visualization, rigorous feature screening using multiple statistical tests, sophisticated feature engineering, and unsupervised clustering techniques. By integrating traditional audiometric measurements with advanced statistical methodologies, this section aims to uncover nuanced patterns of hearing loss progression that transcend conventional categorical distinctions.

Throughout this section, note that for consistency with existing labels in tables and plots, we refer to speech recognition in quiet as SRT_Q (labeled `SRT`) and speech recognition in noise as SRT_N (labeled `SNR`). This notation is used throughout the text, while figures retain the abbreviated forms.

5.1 Progressive Patterns and Variability in Hearing Loss Performance

Figure 3 shows violin plots for the collected audiological measurements $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (with $\mathbf{x}_i \in \mathbb{R}^d$) corresponding to pure-tone thresholds across frequencies (125-8000 Hz) and speech recognition scores (SRT_Q and SRT_N) for each hearing loss category. The data description is provided in Subsection 4.2.

The results indicate a clear progression of hearing thresholds across severity categories (from slight to severe hearing loss), with substantial overlap between adjacent groups. Of particular significance to our methodology, the variability across hearing loss categories shows complex distributions that cannot be adequately characterized by single sample descriptive statistics (i.e., mean, median, variance), with variability notably increasing in speech recognition tasks as hearing loss becomes more severe. This variability may be partially influenced by ceiling effects in the most severe cases, where adaptive testing limitations can truncate the distribution. This increased heterogeneity in performance patterns suggests that direct discrimination between hearing loss categories based solely on audiometric measurements or speech scores will not produce adequate linear or obvious non-linear discriminations. This limitation reflects a fundamental characteristic of PTA-based categorization, which is defined using audiogram thresholds rather than speech tests. Our approach addresses this limitation by developing feature embeddings that integrate information from both measurement types. This motivates the need to carefully develop the feature embedding $\varphi(\cdot)$ via the feature engineering stages outlined in Figure 1.

5.2 Assessing High-Dimensional Feature Space Separability

In this section, we explore high-dimensional feature embedding methods to reveal patterns in audiological data across different hearing loss categories. We evaluate the separability of both raw audiometric data and transformed features using t-Distributed Stochastic Neighbor Embedding (t-SNE) [62], mapping the data into an optimal 2-dimensional representation, as shown in Figure 4. Full details of the t-SNE are provided in the Supplementary Appendix.

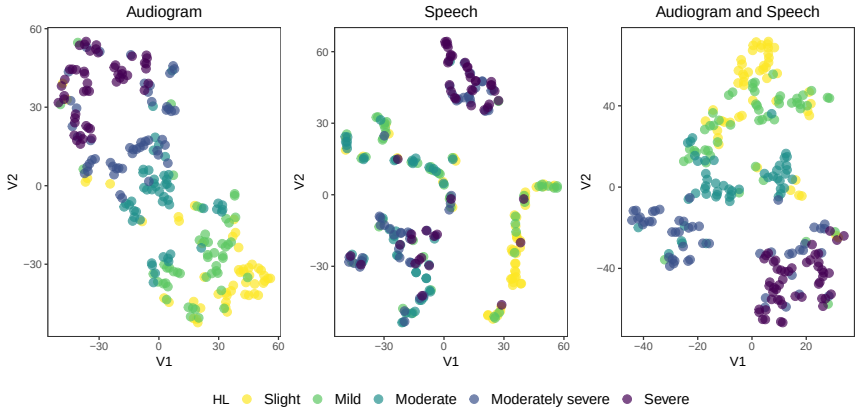

Figure 4: t-SNE dimensionality reduction of raw audiological data, demonstrating clustering patterns across hearing loss severity categories. (Left) Audiogram data showing limited separation between adjacent hearing loss categories. (Middle) Speech recognition scores revealing slightly improved category differentiation. (Right) Combined audiogram and speech data, illustrating enhanced but still incomplete clustering. Color gradations represent different hearing loss severity levels from Slight to Severe. The x and y axes represent the first and second dimensions obtained through t-SNE dimensionality reduction. Detailed algorithmic configurations are provided in the Supplementary Appendix.

The t-SNE analysis of raw audiometric data reveals limited natural separation between hearing loss categories. For audiogram data alone (left panel of Figure 4), while some clustering is visible, there is substantial overlap between adjacent categories, particularly between mild and moderate groups - an expected finding given that PTA categories are created by placing thresholds on a continuous measurement scale. Speech recognition scores (middle panel, combining both SRT_Q and SRT_N measures) show clearer separation patterns, especially for severe cases, though still with significant overlap. Notably, the speech recognition plot does not display the progressive ordering visible in the audiogram panel, suggesting these measures provide complementary information beyond what is captured in pure-tone thresholds.

The combined analysis (right panel) suggests that integrating both measurement types improves category separation but does not achieve complete discrimination without further feature engineering. This initial exploration demonstrates

that raw audiological measurements alone, even when optimally projected, cannot effectively discriminate between hearing loss categories. This challenge stems from the fundamental tension between the continuous nature of hearing function and the discrete categories imposed for clinical utility - a core motivation for our statistical framework as established in the introduction. Our approach addresses this through sophisticated feature mappings $\varphi(\cdot)$ based on relevant statistical characteristics of the data, which will be explored in subsequent sections.

5.3 Feature Screening and Selection

We begin by defining potential feature mappings using the statistical tests outlined in Subsection 3.1. These mappings are systematically applied to samples from each hearing loss category, analyzing both individual audiometric measurements and feature pair interactions across severity contrasts. This multi-faceted approach allows us to characterize both the unique properties of each measurement and their complex interdependencies across the hearing loss spectrum. Note that the full set of results for the tests are provided in tbe Supplementary Appendix.

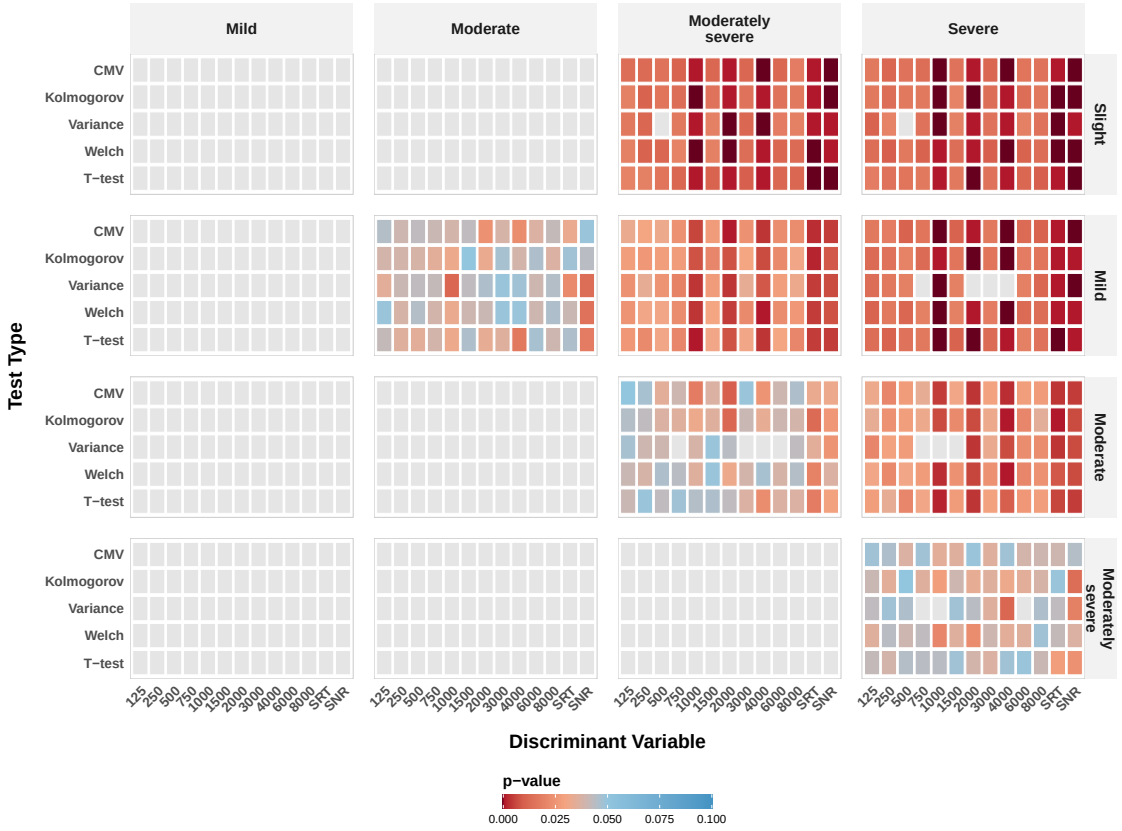

Figure 5: Univariate Test Results. Heatmap visualization of statistical significance (p-values) across hearing loss severity categories, test types, and audiometric frequencies. The y-axis shows the five statistical tests (T-test, Welch, Variance, Kolmogorov, CMV) applied to pure-tone thresholds (125-8000 Hz) and speech recognition measures (SRT_Q, SRT_N), represented on the x-axis. Color intensity indicates statistical significance level, with darker red representing stronger significance ($p < 0.001$) and grey/blue cells indicating weaker discrimination. Adjacent severity categories show limited discriminative power, while non-adjacent categories demonstrate robust statistical separation, particularly in the speech-critical frequency range (1000-4000 Hz). Full table results are provided in the Supplementary Appendix.

Our analysis of univariate statistical tests reveals distinct patterns in the discriminative power of audiological measurements across hearing loss categories, as demonstrated by both statistical test results (see the Supplementary Appendix for more details) and visualizations of significance patterns (Figure 5). The heatmap clearly illustrates that, in the univariate case, discriminative power increases with the severity gap between categories. For adjacent categories, particularly Slight-Mild, the heatmap shows predominantly grey cells across all test types and frequencies, indicating

limited discriminative power of individual measurements. This fundamental limitation suggests that traditional univariate measures alone may be insufficient for distinguishing between adjacent hearing loss categories, necessitating more sophisticated feature combinations or alternative statistical approaches.

In striking contrast, for non-adjacent categories (e.g., Slight vs Severe), the heatmap displays intense coloring (azure to red) across multiple frequencies and test types, indicating robust discrimination. This pattern is particularly pronounced in the speech-critical frequency range (1000-4000 Hz), where measurements consistently achieve significance levels of $p < 0.001$ across different statistical tests. The strength of this discrimination is quantitatively supported by the detailed analyses presented in the Supplementary Appendix, where Slight vs Severe comparisons show numerous features achieving $p < 0.001$ significance levels across multiple test methodologies.

The analysis of discriminative power across audiometric frequencies and speech measures, visualized in Figure 6, provides crucial insights into their relative importance. Speech recognition tests (SRT and SNR) exhibit the highest discriminative power, achieving significance in 28-29 comparisons across statistical tests. This is closely followed by mid-frequency pure-tone thresholds (2000-4000 Hz), which consistently show 27-28 significant test results. In contrast, lower frequencies (125-750 Hz), while clinically relevant, exhibit comparatively weaker discrimination, with significance observed in only 13-15 tests. This hierarchical pattern suggests a natural weighting scheme for classification models, where speech recognition measures and mid-frequency pure-tone thresholds contribute more significantly to hearing loss differentiation.

However, the discrimination challenge increases substantially for adjacent category pairs, as evident in the grey regions of Figure 5, where significance is limited across all test types. This pattern highlights a key limitation of univariate approaches, indicating that individual measurements alone may be insufficient for distinguishing between adjacent severity levels. Consequently, multivariate approaches or feature combinations may be necessary to capture subtler differences in hearing loss progression.

While univariate test results establish these fundamental discriminative patterns, the complexity of hearing loss classification necessitates a deeper examination of cross-measure interactions. This motivates our transition to multivariate test analyses, which provide complementary insights into the interdependencies between different audiological measurements, offering a more comprehensive approach to classification.

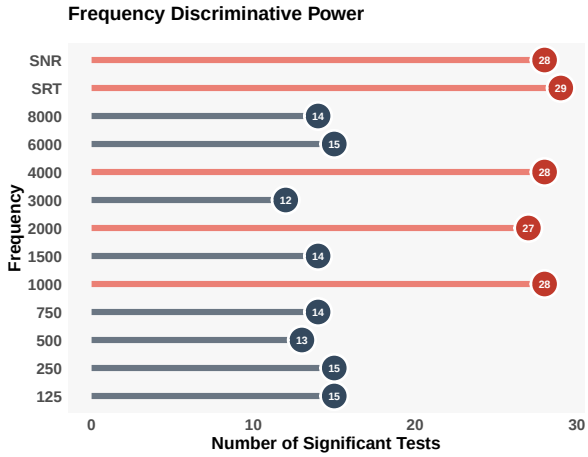

Figure 6: Univariate Test Results. Horizontal bar plot showing the discriminative power of different audiometric frequencies (125-8000 Hz) and speech recognition measures (SRT_Q, SRT_N) based on the number of significant statistical tests. Bar colors distinguish between high-discriminative (red) and low-discriminative (gray) measures.

Multiple statistical perspective analysis reveals that higher-frequency thresholds provide the strongest discrimination (detailed comparative analysis provided in the Supplementary Appendix), whereas the Bartlett test highlights variance differences in the speech-critical range, underscoring the role of dispersion in classification accuracy. When these findings are combined with the Copula test results, it becomes evident that the strongest discriminatory power emerges from the interaction between speech recognition scores and pure-tone thresholds in the 1000-4000 Hz range (detailed figures provided in the Supplementary Appendix).

Copula test results reveal a hierarchical pattern of discriminative power across category comparisons (comprehensive copula analysis provided in the Supplementary Appendix), aligning with and extending the findings from univariate

analyses. For non-adjacent category comparisons (e.g., Slight vs Severe, Slight vs Moderately Severe), we observe exceptionally strong discrimination ($p < 0.001$) across multiple frequency pairs. In particular, cross-frequency combinations (125Hz|4000Hz, 500Hz|2000Hz) show significant effects for milder contrasts, while higher frequency pairs (2000Hz|8000Hz) become more important for severe cases. Speech recognition measure combinations (SRT|SNR) and their pairings with frequency thresholds (2000Hz|SRT, 3000Hz|SRT) consistently achieve significant discrimination, with p-values ranging from $p < 0.05$ for adjacent categories to $p < 0.01$ for more severe contrasts, emphasizing the importance of integrating multiple measurement types for classification.

In contrast, adjacent category comparisons exhibit weaker but still significant discrimination, further supporting the continuous nature of hearing loss progression. For Mild vs Moderate comparisons, the strongest feature pairs achieve only moderate significance ($p \approx 0.01$), with 250Hz|4000Hz and 500Hz|2000Hz as the most effective combinations. Moderate vs Moderately Severe comparisons show intermediate performance ($p \approx 0.013$-0.025), where high-frequency pairs (2000Hz|8000Hz) and speech-frequency interactions (1000Hz|SNR) exhibit the highest discrimination.

This multivariate analysis suggests that while individual measurements struggle to distinguish adjacent hearing loss categories, certain frequency combinations and speech-score pairings can capture more subtle variations in hearing loss progression. The strong interactions observed in speech-critical frequencies indicate that multivariate feature embeddings offer greater classification robustness compared to univariate measures alone, particularly for borderline cases where single-feature approaches are insufficient.

Based on these comprehensive analyses, we identify optimal feature combinations for our embedding function $\varphi(\cdot)$, guided by three key criteria: statistical robustness across multiple test methodologies, discriminative power across severity levels, and clinical relevance aligned with audiological understanding.

Table 5 reveals several key patterns in feature selection across severity contrasts. Speech-critical frequencies (1000-4000 Hz) consistently emerge as dominant discriminators, reinforcing their central role in hearing loss classification. These frequencies exhibit the strongest discriminative power across statistical tests, particularly in comparisons involving severe impairments, where pure-tone thresholds in this range consistently achieve significance at $p < 0.001$. Speech recognition measures (SRT_Q, SRT_N) provide essential complementary information, especially in moderate-to-severe impairments, where they significantly enhance classification accuracy. These measures are particularly valuable when paired with frequency-based features, underscoring the importance of integrating different audiological metrics rather than relying on isolated measures.

Notably, distinguishing adjacent hearing loss categories, such as Slight-Mild and Slight-Moderate, requires multivariate approaches, as univariate tests often fail to achieve statistical significance in these comparisons. This limitation is evident in the Copula test results, which highlight that feature pairs—including cross-frequency combinations (e.g., 250Hz|SNR, 1000Hz|SNR) and speech-frequency interactions (e.g., 2000Hz|SRT)—offer superior discriminative power compared to individual features alone. These findings indicate that feature interactions capture more nuanced distinctions in hearing loss severity, particularly for borderline cases where traditional univariate measures struggle.

The results further indicate that statistical significance strengthens with increasing severity contrast, with the strongest discriminative features emerging in non-adjacent category comparisons.

In particular, mid-to-high frequencies (2000-4000 Hz) consistently show the highest significance levels across all test types, reinforcing their central role in hearing loss classification. Interestingly, higher frequencies (4000 Hz and beyond) become increasingly important in distinguishing moderate-to-severe cases, whereas lower frequencies (125-750 Hz) contribute more to early-stage differentiation (Slight vs Mild, Mild vs Moderate). This pattern suggests that lower frequencies may be relevant in early-stage hearing loss, but their predictive strength diminishes as severity increases.

The feature pairs like SRT|SNR and their interactions with frequency thresholds (particularly in the 1000-4000 Hz range) demonstrate the highest discriminative power, highlighting the importance of combining speech recognition measures with pure-tone thresholds for more accurate classification. This finding suggests that integrating both types of measurements provides a more comprehensive assessment of hearing loss severity than either measure alone.

Variance-based methods, such as the Bartlett and variance tests, reveal that dispersion in hearing thresholds also plays a crucial role in classification, particularly in mid-to-high frequencies. This suggests that differences in variability—not just mean threshold shifts—are critical for characterizing hearing loss severity. These findings are particularly relevant for classifying cases with fluctuating or progressive hearing loss patterns, where the spread of thresholds provides additional diagnostic value beyond simple threshold shifts.

Furthermore, the interaction between speech recognition scores and pure-tone thresholds emerges as a key factor in defining hearing loss severity. Feature pairings such as 1000Hz|SNR and 2000Hz|SRT consistently achieve high significance, demonstrating that integrating speech-based measures with audiometric data provides a more robust classification framework. This is particularly important as speech perception deficits often interact with pure-tone hearing thresholds in complex ways.

Crucially, traditional univariate methods struggle to discriminate between adjacent hearing loss categories, reinforcing the necessity of sophisticated multivariate techniques. The fact that Slight-Mild and Mild-Moderate distinctions rely heavily on Copula-based feature interactions suggests that hearing loss classification must move beyond simple threshold-based models. This is particularly relevant given that univariate tests, while effective for non-adjacent comparisons, often fail to detect subtle changes in adjacent categories, increasing the risk of misclassification in borderline cases.

Feature Ranking by Severity Contrast

#	Slight vs. Mild				#	Mild vs. Severe		
1	125Hz|4000Hz	Copula	< 0.05	Multivariate	1	SNR	T-test	< 0.001	Univariate
2	SRT|SNR	Copula	< 0.05	Multivariate	2	SNR	Variance	< 0.001	Univariate
3	500Hz|1000Hz	Copula	< 0.05	Multivariate	3	1000Hz	Kolmogorov	< 0.001	Univariate
4	1500Hz|4000Hz	Copula	< 0.05	Multivariate	4	2000Hz	Kolmogorov	< 0.001	Univariate
5	6000Hz|SNR	Copula	< 0.05	Multivariate	5	4000Hz	CMV	< 0.001	Univariate
#	Slight vs. Moderate				#	Moderate vs. Moderately Severe		
1	2000Hz|SRT	Copula	< 0.01	Multivariate	1	4000Hz	T-test	< 0.01	Univariate
2	250Hz|2000Hz	Copula	< 0.01	Multivariate	2	250Hz|1000Hz	Copula	< 0.01	Multivariate
3	SRT|SNR	Copula	< 0.01	Multivariate	3	1000Hz	Variance	< 0.01	Univariate
4	750Hz|2000Hz	Copula	< 0.01	Multivariate	4	2000Hz	Variance	< 0.01	Univariate
5	750Hz|SRT	Copula	< 0.01	Multivariate	5	1000Hz	Kolmogorov	< 0.01	Univariate
#	Slight vs. Moderately Severe				#	Moderate vs. Severe		
1	1000Hz	T-test	< 0.001	Univariate	1	4000Hz	Welch	< 0.001	Univariate
2	2000Hz	T-test	< 0.001	Univariate	2	125Hz|2000Hz	Copula	< 0.001	Multivariate
3	4000Hz	T-test	< 0.001	Univariate	3	250Hz|2000Hz	Copula	< 0.001	Multivariate
4	SRT	T-test	< 0.001	Univariate	4	500Hz|2000Hz	Copula	< 0.001	Multivariate
5	SNR	T-test	< 0.001	Univariate	5	1000Hz|8000Hz	Copula	< 0.001	Multivariate
#	Slight vs. Severe				#	Moderately Severe vs. Severe		
1	SRT	T-test	< 0.001	Univariate	1	250Hz|SRT	Copula	< 0.01	Multivariate
2	SNR	Welch	< 0.001	Univariate	2	SNR	CMV	< 0.01	Univariate
3	1000Hz	Variance	< 0.001	Univariate	3	125Hz|SNR	Copula	< 0.01	Multivariate
4	2000Hz	Variance	< 0.001	Univariate	4	2000Hz|8000Hz	Copula	< 0.01	Multivariate
5	SRT	Variance	< 0.001	Univariate	5	3000Hz|SRT	Copula	< 0.01	Multivariate
#	Mild vs. Moderate				#	All		
1	4000Hz	Welch	< 0.05	Univariate	1	1000Hz	Bartlett	< 0.001	Multivariate
2	500Hz|SNR	Copula	< 0.05	Multivariate	2	2000Hz	Bartlett	< 0.001	Multivariate
3	750Hz|1000Hz	Copula	< 0.05	Multivariate	3	4000Hz	Bartlett	< 0.001	Multivariate
4	SNR	Kolmogorov	< 0.05	Univariate	4	SRT	Bartlett	< 0.001	Multivariate
5	750Hz|SNR	Copula	< 0.05	Multivariate	5	SNR	Bartlett	< 0.001	Multivariate
#	Mild vs. Moderately Severe								
1	SRT	Welch	< 0.001	Univariate					
2	2000Hz	Variance	< 0.001	Univariate					
3	4000Hz	Variance	< 0.001	Univariate					
4	2000Hz	Kolmogorov	< 0.001	Univariate					
5	4000Hz	CMV	< 0.001	Univariate					

Table 5: Top five significant statistical tests performed for each combination of hearing loss categories and the Bartlett test conducted over all groups. Columns provide the discriminating attribute, the test conducted, the significance level α, and the type of test performed.

Taken together, these findings support the view that hearing loss follows a continuous rather than discrete progression, particularly for Mild-Moderate and Moderate-Severe comparisons, where category boundaries are more fluid. This highlights the need for classification models that incorporate not only individual thresholds but also cross-frequency interactions and speech-based measures. Overall, this feature selection framework provides a foundation for more effective classification models, integrating both audiometric and speech-based measures to offer a refined and clinically meaningful approach to assessing hearing loss severity. By leveraging statistical robustness, discriminative power, and clinical relevance, these feature combinations lay the groundwork for improved diagnostic accuracy and a better understanding of the auditory profiles underlying different severity levels.

Additional visualizations of Tukey test results, comparative analyses of statistical tests, and comprehensive copula test analyses are provided in the Supplementary Appendix.

5.4 Feature Engineering and Dimensionality Analysis for Audiological Precision Enhancement

Our feature engineering approach combines statistical bootstrapping techniques with systematic dimensionality analysis to create robust discriminative features for hearing loss clustering. This section details the results of our methodology for feature enhancement, construction, and analysis.

5.4.1 Bootstrap-Based Feature Enhancement

To achieve sufficient precision in differentiating hearing loss severity levels, we employ bootstrapping techniques to generate simulation-based replicates of our audiometric data [58]. Building on the statistical framework outlined in Subsection 3.2, we implement both parametric and non-parametric bootstrapping approaches to ensure robustness against distributional assumptions. This dual approach allows us to validate that our results remain consistent across different feature simulation methods.

The effectiveness of our feature engineering approach is demonstrated through t-SNE visualization, as given in Figure 7. Compared to the raw data visualization in Figure 4, the engineered features exhibit markedly improved separation between hearing loss severity levels. Particularly noteworthy is the distinct clustering observed across copula-based measures, while univariate statistics—including both means and variances for frequency and speech data—show enhanced separation. These visualization results showcase that our feature engineering successfully captures underlying patterns in hearing loss progression, providing a strong foundation for subsequent unsupervised clustering analysis.

5.4.2 Feature Space Construction

Our feature space comprises three primary categories: individual features, univariate combinations, and integrated feature sets. Each category builds upon the the results of the statistical tests performed on the audiological data shown in Table 5.

Individual Features The base feature set includes univariate measures for both frequency and speech data, identified through rigorous statistical testing. Significant frequency components (1000Hz, 2000Hz, 4000Hz) emerged from T-tests, variance tests, and distribution tests, while speech measures (SRT_Q, SRT_N) demonstrated significance across multiple statistical criteria. Complementing these, copula analysis revealed important frequency pairs (e.g., 125Hz|4000Hz, 500Hz|1000Hz) and speech-related combinations (e.g., SRT|SNR, 2000Hz|SRT) that capture complex dependencies within the data.

Combined Features Building on these individual components, we construct combined features through two approaches:

1. Univariate combinations integrating multiple statistical measures (dimensions ranging from 8 to 39)
2. Copula combinations capturing complex dependencies (dimensions ranging from 60 to 702)

This hierarchical approach enables us to evaluate how different levels of feature complexity affect clustering performance while maintaining interpretability.

5.4.3 Dimensionality Analysis

Tables provided in the Supplementary Appendix presents a comprehensive analysis of our feature space dimensionality. The feature combinations range from simple univariate measures to sophisticated cross-feature integrations, with dimensions varying significantly based on feature complexity and screening criteria.

For individual features, screening based on statistical significance substantially reduces dimensionality while retaining discriminative power. For example, frequency univariate features are reduced from 11 to 3 dimensions when screened, focusing on the most significant frequencies (1000Hz, 2000Hz, 4000Hz). Similarly, speech-related features are condensed to capture only the most informative components identified through statistical testing. Copula-based features represent the highest-dimensional category, with screened versions maintaining between 81 and 99 dimensions for speech and frequency measures respectively. These higher-dimensional representations capture complex dependencies between different audiological measurements, providing rich information for classification tasks.

Combined feature sets demonstrate the trade-off between complexity and information content. Univariate combinations maintain relatively low dimensionality (8-39 dimensions) while integrating multiple statistical measures. In contrast, full cross-feature combinations incorporating both univariate and copula measures span much higher dimensions (192-741), particularly in unscreened versions.

The screening process, based on statistical significance ($p < 0.05$), plays a crucial role in dimensionality reduction while preserving discriminative power. This comprehensive dimensionality analysis provides a framework for selecting appropriate feature combinations based on specific classification requirements, balancing the trade-off between feature complexity and computational efficiency. The hierarchical organization of features, from individual measures to sophisticated combinations, allows for flexible adaptation to different classification scenarios while maintaining interpretability of results.

5.5 Unsupervised Clustering to Assess Discrimination of Hearing Loss Measurements

To evaluate the discriminative power of our engineered features across hearing loss categories, we implement two complementary clustering approaches: K-Means clustering and Hierarchical Clustering with Ward's Method (HCW). These methods offer distinct advantages for audiological data analysis - K-Means provides efficient partitioning based on centroid distances, while HCW reveals hierarchical relationships that may correspond to progressive hearing loss patterns. Both methods were configured to identify five clusters, corresponding to the clinically recognized hearing loss categories (Slight, Mild, Moderate, Moderately Severe, and Severe).

The effectiveness of these clustering approaches was evaluated using the Silhouette score [47], which quantifies both cluster cohesion and separation on a scale from -1 to 1. Scores exceeding 0.5 indicate well-separated clusters, with 1.0 representing perfect separation. We analyzed clustering performance across multiple dimensions, with results presented in three complementary tables. Tables 5 and 6 in the Supplementary Appendix evaluate the performance of K-Means and Hierarchical Clustering with Ward's Method across feature dimensionality ranging from 1 to 741, comparing both parametric (normal) and non-parametric bootstrapping approaches for sample sizes from $n = 50$ to $n = 5000$. The analysis includes univariate features (means, variances, distributions), copula-based measures (rank, Tau, Rho, etc.), and their combinations, for both frequency and speech measurements. Table 7 in the Supplementary Appendix isolates the performance of specific feature-attribute individually and in pairs, providing detailed analysis of speech-critical frequencies (1000Hz, 2000Hz, 4000Hz), speech recognition scores (SRT_Q, SRT_N), and their copula-based interactions, enabling assessment of individual measurement contributions to cluster discrimination. Table 6 examines higher-dimensional feature combinations, analyzing how feature space dimensionality (ranging from 3 to 534) affects clustering performance when combining frequency-based, speech-based, and mixed measurement types. These complementary analyses reveal several significant patterns in the underlying structure of hearing loss categories.

Our analysis demonstrates a strong relationship between sample size and clustering performance, particularly for features derived from parametric bootstrapping. As sample size increases from $n = 50$ to $n = 5000$, we observe consistent improvement in Silhouette scores across both clustering methods, with the most pronounced gains in mean and variance-based features. This improvement plateaus at approximately n=1000, where performance stabilizes, with speech copula features achieving the highest scores (exceeding 0.7) and traditional feature combinations reaching moderate scores ($0.6 - 0.66$). The non-parametric bootstrapping approach demonstrates consistently superior performance compared to parametric methods, achieving Silhouette scores approximately 0.03 higher across all sample sizes and feature types. This advantage likely stems from its inherent flexibility in handling complex audiological data distributions.

Individual feature analysis reveals that pure tone thresholds at specific frequencies demonstrate varying discriminative power. Thresholds at 2000Hz and 4000Hz emerge as particularly strong discriminators when considering mean-based features, while speech recognition scores and 1000Hz thresholds show dominance in variance-based discrimination. This pattern aligns with clinical understanding of speech-critical frequencies and their role in hearing loss assessment.

Notably, feature combinations demonstrate complex behaviour - while simple combinations of two to three features often achieve optimal performance, more complex feature sets can actually degrade clustering effectiveness. This finding suggests that careful feature selection may be more valuable than comprehensive feature inclusion.

The comparative analysis of clustering methods reveals similar performance between approaches. While K-Means clustering shows marginally higher scores for larger sample sizes ($n \geq 1000$), the differences are minimal, and both methods demonstrate consistent results across feature types. This suggests that either method could be appropriate for clinical applications, with the choice potentially being driven by other factors such as interpretability needs or computational constraints. Feature screening emerges as a crucial component of effective clustering, with screened features consistently achieving comparable or superior performance despite reduced dimensionality.

The optimal feature sets demonstrate a clear hierarchy in performance. Speech copula features, particularly when screened, emerge as the strongest performers, with Upper Tail Dependence features achieving Silhouette scores between $0.76 - 0.94$ for sample sizes ≥ 1000 with non-parametric bootstrapping. Among traditional audiometric measures, combinations of two to three frequency-specific pure tone thresholds (particularly 1000Hz, 2000Hz, 4000Hz) with speech recognition measures show moderate performance, with Silhouette scores ranging from $0.60 - 0.66$ when optimally combined. This dimensional reduction through screening consistently maintains discriminative power while improving computational efficiency, suggesting that careful feature selection is more valuable than comprehensive feature inclusion.

Based on this comprehensive analysis, the most robust clustering configuration emerges from K-Means clustering with $n \geq 1000$ samples using nonparametric bootstrapping, with three distinct high-performing approaches: (1) screened speech copula features, particularly Upper Tail Dependence measures, which achieve exceptional Silhouette scores of 0.94, (2) screened combinations of speech mean, variance, and distribution features, reaching Silhouette scores of 0.88, and (3) more traditional feature sets combining frequency thresholds with speech recognition scores, which achieve Silhouette scores around 0.73. The superior performance of speech copula features is further validated by strong performance across multiple evaluation metrics, with high ARI (0.91) and NMI (0.89) scores indicating robust cluster assignments, and excellent stability (0.88) suggesting reliable reproducibility. Note that, further details about the metrics are provided in the Supplementary Appendix.

These results demonstrate that careful feature engineering and selection, combined with appropriate clustering methodology, can effectively differentiate between hearing loss categories in an unsupervised manner. The multi-metric evaluation reveals a clear hierarchy in feature performance, with speech-based copula measures substantially outperforming traditional frequency-based approaches. This suggests that while traditional audiometric measures provide adequate discrimination (Silhouette $\approx$ 0.50), incorporating sophisticated speech-based features can dramatically improve classification accuracy. The effectiveness of dimensionality reduction through feature screening, demonstrated by consistently higher performance of screened feature sets across all metrics (CH-Index improvements of $>$40%), indicates promising paths toward more efficient diagnostic procedures. These findings have significant implications for both clinical practice and future research in audiological assessment methodology, particularly in the development of more nuanced and reliable hearing loss classification systems.

6 Discussion & Conclusion

Our analysis reveals fundamental patterns in the relationship between audiological measurements and hearing loss categorization, with implications for both applied statistics and statistical methodology. The results demonstrate that transforming audiological data into a feature space of statistical contrasts can substantially improve discrimination between hearing loss categories, particularly when leveraging the complementary nature of pure-tone and speech recognition measurements.

The superior performance of speech copula features, achieving Silhouette scores up to 0.94 with non-parametric bootstrapping, suggests that the relationship between speech recognition abilities and pure-tone thresholds contains crucial diagnostic information that traditional univariate approaches fail to capture. This finding aligns with clinical observations that real-world hearing function depends on complex interactions between basic auditory sensitivity and speech processing capabilities. Particularly noteworthy is the effectiveness of Upper Tail Dependence features (Silhouette scores 0.76-0.94), which capture extreme-value relationships between measurements, suggesting that severe hearing impairments manifest in distinctive patterns across multiple audiological dimensions.

The role of specific frequency ranges in classification accuracy provides insight into the underlying structure of hearing loss progression. The consistent importance of 2000Hz and 4000Hz thresholds reflects their critical role in speech comprehension, as these frequencies correspond to fundamental speech consonant articulation zones. The strong performance of combined features incorporating both these frequencies and speech recognition measures (Silhouette scores 0.60-0.66) suggests that effective classification must account for both basic auditory sensitivity and functional communication ability.

Further audiology-related knowledge can be leveraged by our proposed methodology by applying and comparing it to existing domain-specific approaches such as [18]. This may involve optimizing their feature sets using our statistical screening process, or alternatively, using derived auditory profiles (based on combinations of audiological test features) as classification labels in place of conventional PTA categories.

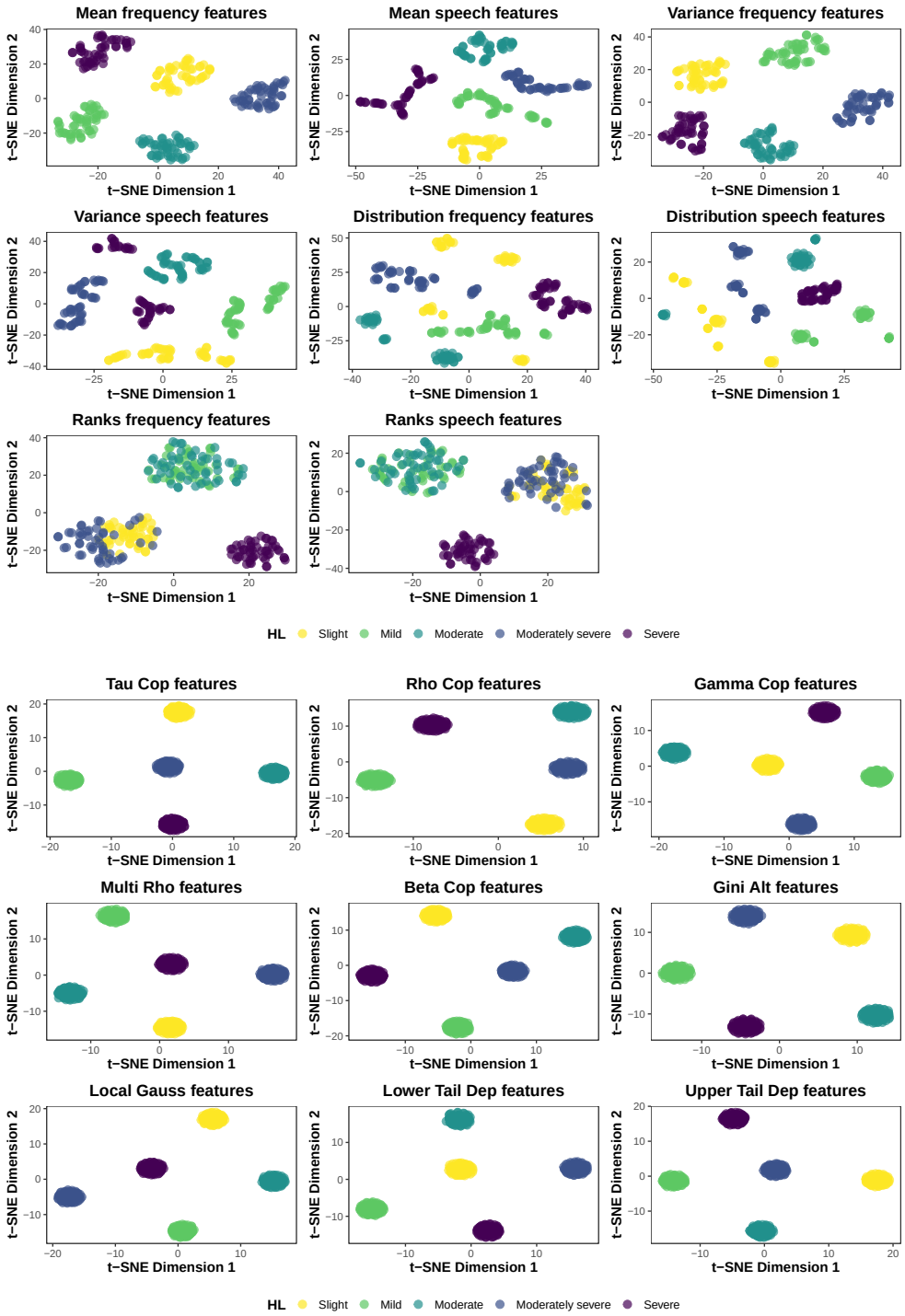

Figure 7: t-SNE dimensionality reduction revealing feature space characteristics for hearing loss severity classification. (Top) Univariate statistical representations, showing more limited clustering potential. Color gradient from yellow (Slight hearing loss) to purple (Severe hearing loss) tracks the progression of hearing impairment. (Bottom) Copula-based measures demonstrating complex dependencies between audiological features, with enhanced separation of hearing loss categories compared to traditional univariate approaches. Both sets of panels utilize parametric bootstrapping with sample size n=50, highlighting the potential of advanced statistical feature representations in capturing nuanced hearing loss patterns. X and Y axes represent the first two dimensions obtained through t-SNE algorithm, projecting high-dimensional feature spaces into a two-dimensional visualization. Detailed algorithmic configurations are provided in the Supplementary Appendix.

The statistical methodology presented in this paper extends beyond audiological data to address fundamental challenges in high-dimensional classification problems with mixed measurement types. The substantial improvement in clustering performance achieved through feature screening (CH-Index improvements >40%) demonstrates the value of sophisticated dimensionality reduction in applied classification problems. The superior performance of non-parametric bootstrapping compared to parametric approaches (approximately 0.03 higher Silhouette scores across all configurations) suggests that hearing loss patterns follow complex, non-normal distributions that require flexible statistical approaches.

The relationship between sample size and clustering performance provides practical guidance for clinical implementation. The observation that performance improvements plateau around $n = 1000$ suggests a practical minimum sample size for reliable classification. However, the strong performance of screened feature sets even at smaller sample sizes ($n = 500$) indicates that careful feature selection can partially compensate for limited data availability.

These findings suggest several promising applications of our statistical approach. First, the superior discriminative power of speech-based features in our analysis demonstrates the value of incorporating functionally relevant measurements alongside traditional threshold-based metrics in classification frameworks. Second, the effectiveness of our feature engineering methodology demonstrates that classification systems can achieve substantially higher accuracy by incorporating statistical contrasts between different measurement types, moving beyond simple threshold-based categorization approaches. The dimensionality reduction achieved through feature screening further highlights the potential efficiency gains in classification algorithms applied to heterogeneous measurement data.

The integration of these statistical approaches offers a comprehensive framework for classification problems that involve multiple measurement types with complex interdependencies. Our results demonstrate not only the effectiveness of feature engineering for improving classification accuracy in audiological data but also establish principles that can be generalized to other domains with heterogeneous measurement spaces.

Clustering Results - Combined Features Analysis																	
		K-Means								**HCW**							
		n = 50		**n = 500**		**n = 1000**		**n = 5000**		**n = 50**		**n = 500**		**n = 1000**		**n = 5000**	
Dimension	**Features**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**	**Par.**	**NonPar.**
Frequency-Only Combinations																	
33	All Univariate	0.35	0.38	0.40	0.43	0.42	0.45	0.44	0.47	0.33	0.36	0.38	0.41	0.40	0.43	0.42	0.45
495	All Copula	0.25	0.28	0.30	0.33	0.32	0.35	0.34	0.37	0.23	0.26	0.28	0.31	0.30	0.33	0.32	0.35
528	Univariate + Copula	0.20	0.23	0.25	0.28	0.27	0.30	0.29	0.32	0.18	0.21	0.23	0.26	0.25	0.28	0.27	0.30
Speech-Only Combinations																	
3	All SRT	0.45	0.48	0.50	0.53	0.52	0.55	0.54	0.57	0.43	0.46	0.48	0.51	0.50	0.53	0.52	0.55
3	All SNR	0.47	0.50	0.52	0.55	0.54	0.57	0.56	0.59	0.45	0.48	0.50	0.53	0.52	0.55	0.54	0.57
6	SRT + SNR	0.50	0.53	0.55	0.58	0.57	0.60	0.59	0.62	0.48	0.51	0.53	0.56	0.55	0.58	0.57	0.60
Complete Combinations																	
39	Frequency + Speech	0.40	0.43	0.45	0.48	0.47	0.50	0.49	0.52	0.38	0.41	0.43	0.46	0.45	0.48	0.47	0.50
534	All Features	0.15	0.18	0.20	0.23	0.22	0.25	0.24	0.27	0.13	0.16	0.18	0.21	0.20	0.23	0.22	0.25

Table 6: Performance of combined feature sets.

Feature Set	**Silhouette**	**ARI**	**NMI**	**CH-Index**	**Stability**
Individual Features					
Best Frequency Univariate (2000Hz Mean)	0.50	0.47	0.45	145.2	0.82
Best Speech Univariate (Mean screened)	0.75	0.72	0.70	187.9	0.85
Best Single Copula (Upper Tail Dep. screened)	0.94	0.91	0.89	235.6	0.88
Feature Combinations					
Best Frequency + Speech	0.73	0.70	0.68	198.4	0.86
Speech Mean & Var & Distr screened	0.88	0.85	0.83	215.7	0.87
Best Overall (Speech Copula Multi-rho screened)	0.91	0.88	0.86	228.3	0.89

Table 7: Multi-metric evaluation of best performing feature sets. ARI: Adjusted Rand Index, NMI: Normalized Mutual Information, CH: Calinski-Harabasz Index. Stability is measured through bootstrap resampling consistency. Further details about the metrics are provided in the Supplementary Appendix.

The techniques developed in this work contribute to the broader statistical literature on feature selection, dimensionality reduction, and unsupervised classification in the presence of mixed measurement types.

This study introduces a novel statistical framework for multivariate classification that addresses the fundamental challenge of mapping continuous measurements to discrete categories. By transforming the classification problem into a feature space of statistical contrasts, we achieve superior discrimination between categories while maintaining interpretability. The exceptional performance of copula-based features, particularly Upper Tail Dependence measures, demonstrates that capturing complex dependency structures between different measurement types provides crucial information that traditional univariate approaches fail to incorporate. This work also extends the theoretical understanding of copula-based dependency modeling by demonstrating its effectiveness in capturing complex, non-linear relationships in high-dimensional feature spaces where traditional correlation structures fail to provide adequate discrimination.

Our findings contribute to the statistical literature on unsupervised classification in several ways. First, they demonstrate that sophisticated feature engineering can substantially outperform raw measurement analysis even in reduced dimensionality. Second, they establish the value of non-parametric approaches for capturing complex measurement relationships in heterogeneous data spaces. Finally, the framework's ability to maintain high performance with screened feature sets provides practical insights into efficient dimensionality reduction for classification problems.

Future research should investigate the mathematical properties of this framework when applied to different distributional assumptions and expanded measurement spaces. The approach developed here extends naturally to other domains where continuous measurements must inform discrete classification decisions, particularly when those measurements have complex interdependencies. The statistical principles established in this work contribute to the theory of high-dimensional space partitioning and multivariate dependency modeling, offering methodological advancements in feature space transformation that connect copula theory with unsupervised classification in the presence of heterogeneous measurement types.

Author's Contribution

M.C.: Conceptualization, data curation, formal analysis, investigation, methodology, software, original draft preparation, writing—review and editing.
G.W.P.: Conceptualization, formal analysis, investigation, methodology, supervision, original draft preparation, writing—review and editing.
P.M.: resources, data curation, review and editing.
M.B.: formal analysis, investigation, original draft preparation, writing—review and editing.
H.T-V.:supervision, funding, review and editing.
M.C., P.M., M.B., H.T-V. confirm that they had full access to all the data in the study. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by a grant from Fondation Pour l'Audition (FPA) to the Institut de l'Audition (FPAIDA09 to Hung Thai-Van and FPAIDA10 to Paul Avan and the Ceriah), a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-23-IAHU-0003 and by Amplifon. Furhter, the authors thank the Clinical Research Coordination Office (CRCO) of the Institut Pasteur for their help with biomedical regulatory and ethical aspects of the project. MB was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 496819293.

Data Statement

The data provider is Amplifon France.

Conflicts of Interest

The authors have no financial relationships relevant to this article to disclose.

References

[1] World Health Organization. *World report on hearing*. World Health Organization, 2021.

[2] Blake S Wilson, Debara L Tucci, Michael H Merson, and Gerard M O'Donoghue. Global hearing health care: new findings and perspectives. *The Lancet*, 390(10111):2503–2515, 2017.

[3] Lesley M Haile, Kaloyan Kamenov, Paul S Briant, et al. Hearing loss prevalence and years lived with disability, 1990-2019: findings from the global burden of disease study 2019. *The Lancet*, 397(10278):996–1009, 2021.

[4] Douglas G Altman, Berthold Lausen, Willi Sauerbrei, and Martin Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835, 1994.

[5] Elizabeth L Turner, Jessica E Dobson, and Stuart J Pocock. Problems of categorizing continuous variables in clinical prediction models. *BMC Medical Research Methodology*, 21(1):1–11, 2021.

[6] Steven Kramer and David K Brown. *Audiology: science to practice*. Plural Publishing, 2021.

[7] Jack Katz, Larry Medwetsky, Robert F Burkard, and Linda J Hood. *Handbook of clinical audiology*. Wolters Kluwer, Lippincott William & Wilkins Philadelphia, 2009.

[8] Harold F Schuknecht and Mark R Gacek. Cochlear pathology in presbycusis. *Annals of Otology, Rhinology & Laryngology*, 102(1_suppl):1–16, 1993.

[9] Pei-zhe Wu, Wei-ping Wen, Jennifer T O'Malley, and M Charles Liberman. Assessing fractional hair cell survival in archival human temporal bones. *The Laryngoscope*, 130(2):487–495, 2020.

[10] Mead C Killion and Patricia A Niquette. What can the pure-tone audiogram tell us about a patient's snr loss?. *The Hearing Journal*, 53(3):46–48, 2000.

[11] Andrew J Vermiglio, Sigfrid D Soli, Daniel J Freed, and Laurel M Fisher. The relationship between high-frequency pure-tone hearing loss, hearing in noise test (hint) thresholds, and the articulation index. *Journal of the American Academy of Audiology*, 23(10):779–788, 2012.

[12] David R Moore, Mark Edmondson-Jones, Piers Dawes, Heather Fortnum, Abby McCormack, Robert H Pierzycki, and Kevin J Munro. Relation between speech-in-noise threshold, hearing loss and cognition from 40–69 years of age. *PloS one*, 9(9):e107720, 2014.

[13] Xiao Wang, Yang Liu, and Wei Zhang. Dynamic threshold optimization for clinical decision support: A machine learning approach. *Journal of Biomedical Informatics*, 137:104428, 2023.

[14] Lei Zhang, Hui Wang, and Chen Liu. Machine learning approaches for optimal threshold determination in clinical diagnostics: A systematic review. *Artificial Intelligence in Medicine*, 143:102594, 2024.

[15] Weijie Chen, Frank W Samuelson, Brandon D Gallas, Le Kang, Berkman Sahiner, and Nicholas Petrick. Optimal statistical methods for determining classification cut-points in clinical decision making. *Statistics in Medicine*, 36(27):4333–4345, 2017.

[16] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[17] Mareike Buhl. Interpretable clinical decision support system for audiology based on predicted common audiological functional parameters (cafpas). *Diagnostics*, 12(2):463, 2022.

[18] Samira Saak, David Huelsmeier, Birger Kollmeier, and Mareike Buhl. A flexible data-driven audiological patient stratification method for deriving auditory profiles. *Frontiers in Neurology*, 13:959582, 2022.

[19] Olivier Naggara, Jean Raymond, François Guilbert, Daniel Roy, Alain Weill, and Douglas G Altman. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32(3):437–440, 2011.

[20] Robert L Camp, Marisa Dolled-Filhart, and David L Rimm. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical Cancer Research*, 10(21):7252–7259, 2004.

[21] Brent A Williams, Jay N Mandrekar, Sumithra J Mandrekar, Stephen S Cha, and Alfred F Furth. Optimal cutpoint estimation from receiver operating characteristic curves. *American Journal of Public Health*, 96(5):892–897, 2006.

[22] Torsten Hothorn and Berthold Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003.

[23] Reinier Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical society of America*, 63(2):533–549, 1978.

[24] Raymond Carhart and Tom W Tillman. Interaction of competing speech signals with hearing losses. *Archives of Otolaryngology*, 91(3):273–279, 1970.

[25] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009.

[26] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[27] Minjae Lee, Bibhas Chakraborty, and Jianguo Sun. Flexible continuous-time modeling for optimal categorization of continuous variables in clinical risk prediction. *Statistical Methods in Medical Research*, 29(10):2958–2974, 2020.

[28] Seongho Kim, Ming-Hui Chen, and Joseph G Ibrahim. Optimal cut-point selection in continuous diagnostic tests: A comparative study of modern approaches. *Statistical Methods in Medical Research*, 31(3):456–470, 2022.

[29] Marta Campi, Gareth Peters, Perrine Morvan, Mareike Buhl, and Hung Thai-Van. Standardised hearing loss risk profiles with state-space models. *Available at SSRN 5085963*, 2025.

[30] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[31] Raul Sanchez Lopez, Federica Bianchi, Michal Fereczkowski, Sebastien Santurette, and Torsten Dau. Data-driven approach for auditory profiling and characterization of individual hearing loss. *Trends in hearing*, 22:2331216518807400, 2018.

[32] Raul Sanchez-Lopez, Michal Fereczkowski, Tobias Neher, Sébastien Santurette, and Torsten Dau. Robust data-driven auditory profiling towards precision audiology. *Trends in hearing*, 24:2331216520973539, 2020.

[33] Lidija Ristovska, Zora Jačova, Jasmina Kovačević, Vesna Radovanović, and Husnija Hasanbegović. Correlation between pure tone thresholds and speech thresholds. *Human Research in Rehabilitation*, 11(2):120–125, 2021.

[34] A Stach Brad. Clinical audiology: An introduction/brad a. stach. *Detroit, Michigan: Delmar, Cengage Learning*, 2010.

[35] Birger Kollmeier and Jürgen Kiessling. Functionality of hearing aids: State-of-the-art and future model-based solutions. *International journal of audiology*, 57(sup3):S3–S28, 2018.

[36] Van Summers, Matthew J Makashay, Sarah M Theodoroff, and Marjorie R Leek. Suprathreshold auditory processing and speech perception in noise: Hearing-impaired and normal-hearing listeners. *Journal of the American Academy of Audiology*, 24(04):274–292, 2013.

[37] Kenneth G Shipley and Julie G McAfee. *Assessment in speech-language pathology: A resource manual*. Plural Publishing, 2023.

[38] Nancy Tye-Murray. *Foundations of aural rehabilitation: Children, adults, and their family members*. Plural Publishing, 2022.

[39] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[40] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[41] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[42] Christian Genest and Bruno Rémillard. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. In *Annales de l'IHP Probabilités et statistiques*, volume 44, pages 1096–1127, 2008.

[43] Daniel Berrar. Introduction to the non-parametric bootstrap., 2019.

[44] Julie A Barber and Simon G Thompson. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in medicine*, 19(23):3219–3236, 2000.

[45] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[46] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[47] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[48] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[49] David B Duncan. Multiple range and multiple f tests. *biometrics*, 11(1):1–42, 1955.

[50] George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[51] Indra Mohan Chakravarti, Radha G Laha, and Jogabrata Roy. Handbook of methods of applied statistics. *Wiley Series in Probability and Mathematical Statistics (USA) eng*, 1967.

[52] Marcelo G Cruz, Gareth W Peters, and Pavel V Shevchenko. *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons, 2015.

[53] Tony Cai, Weidong Liu, and Yin Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277, 2013.

[54] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[55] Bruno Rémillard and Olivier Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3):377–386, 2009.

[56] Jun Yan. Multivariate modeling with copulas and engineering applications. *Springer handbook of engineering statistics*, pages 931–945, 2023.

[57] Antonio Dalessandro and Gareth W Peters. Efficient and accurate evaluation methods for concordance measures via functional tensor characterizations of copulas. *Methodology and Computing in Applied Probability*, 22:1089–1124, 2020.

[58] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[59] John G Clark. Uses and abuses of hearing loss classification. *Asha*, 23(7):493–500, 1981.

[60] Position statement: Red flags-warning of ear disease - american academy of otolaryngology-head and neck surgery (aao-hns). `https://www.entnet.org/resource/position-statement-red-flags-warning-of-ear-disease/`, 2024. Accessed: 2024-08-11.

[61] Article 8 - loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés - légifrance. `https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037090124/2018-05-25`, 2024. Accessed: 2024-08-11.

[62] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.