WIDEBAND ABSORBANCE ANALYSIS: APPLYING MACHINE LEARNING TO DETECT CONDUCTIVE HEARING LOSS

Introduction

Conductive hearing loss is a prevalent health concern in infants and young children, as it can negatively impact communication and overall well-being. Timely detection and intervention are crucial for successful treatment while untreated cases may lead to more severe ear conditions in the long term.

Recent advancements in audiological technology have introduced wideband absorbance (WBA) tests, providing detailed absorbance levels across multiple frequencies. However, the complexity of the output poses challenges for audiologists to analyze in clinical settings.

In this project, we employ statistical methods and machine learning techniques to identify influential frequencies for accurate hearing loss detection, in order to streamline the examination of the WBA chart.

Methodology

The data contains 239 wideband absorbance test results of children's ears, with absorbance levels measured across 107 frequencies ranging from 226 Hz to 8000 Hz. Notably, the dataset exhibits class imbalance, with only 38 observations diagnosed with conductive hearing loss, while the remaining 201 samples represent healthy ears.

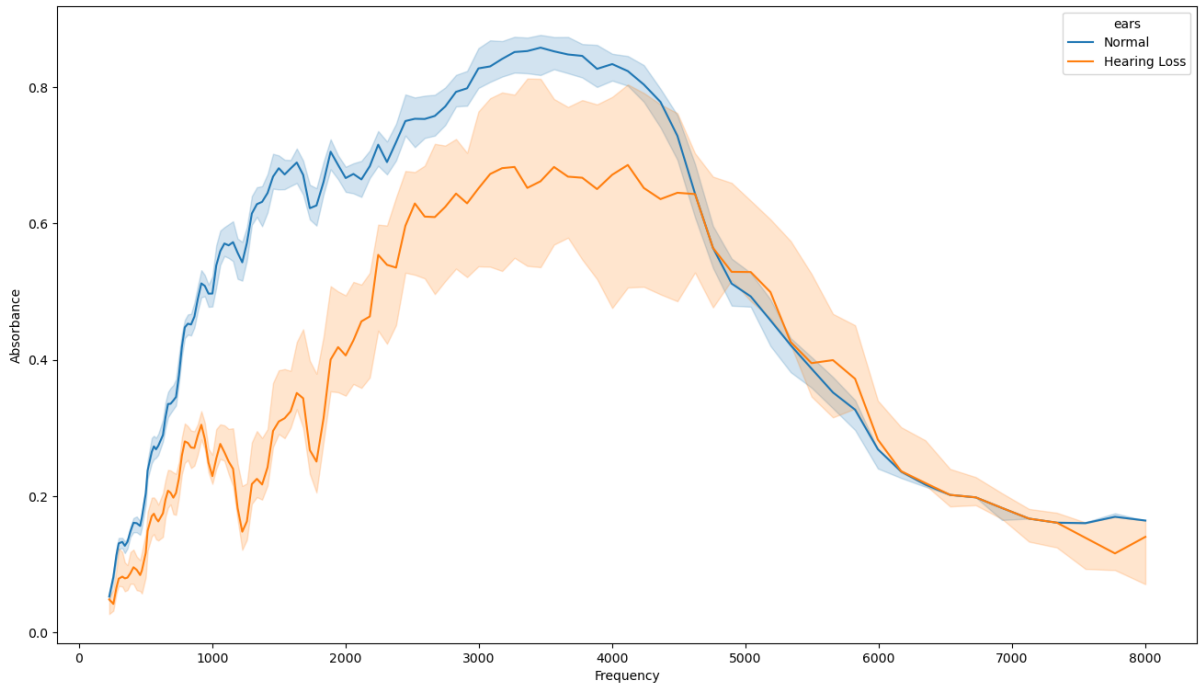
To address this imbalance set, we will employ oversampling techniques, specifically SMOTE, SMOTEENN, and SMOTETomek. These methods aim to augment the minority class, creating a more balanced training set in hopes to enhance accuracy and performance.

For selecting key frequencies, we will utilize embedded feature selection methods, applying Lasso (L1) and Ridge (L2) regularizations to filter out less influential frequencies, focusing on those that contribute significantly to the classification task.

Finally, we opted for the Support Vector Machines (SVM) model due to its suitability for high-dimensional data, given that our analysis encompasses 107 frequencies as features. In particular, we will use the linear kernel for our SVM to ensure transparent and interpretable classifications.This choice allows for a clearer understanding of the WBA chart and aids in gaining valuable insights from the model's predictions.

Results

1. EDA



The line chart above provides an initial insight into the data, suggesting a distinct separation in absorbance levels between frequencies ranging from 1000 to 2000 Hz. In more detail, normal ears exhibit consistently higher absorbance levels in the range of 0.6 to 0.8, while hearing loss ears display lower absorbance levels (<0.4) in those specific frequency regions.

2. Models

Model	Data Sample	Feature Selection	Selected Frequencies	Training Recall	Validation Recall
1	original	Lasso (L1)	1296, 3886, 4117, 5656, 5822	82.88	81.25
2	original	Ridge (L2)	1296, 1334, 3886, 5656, 5822	86.30	78.75
3	smote	Lasso (L1)	727, 1373, 4000, 5187, 8000	85.80	83.75
4	smote	Ridge (L2)	1334, 4117, 5495, 5656, 5822	87.55	82.50
5	smoteenn	Lasso (L1)	890, 1029, 1414, 1587, 5993	85.05	88.75

6	smoteenn	Ridge (L2)	257, 280, 1542, 1587, 5993	85.89	83.75
7	smotetomek	Lasso (L1)	727, 1373, 4000, 5187, 8000	85.80	83.75
8	smotetomek	Ridge (L2)	1334, 4117, 5495, 5656, 5822	87.55	82.50

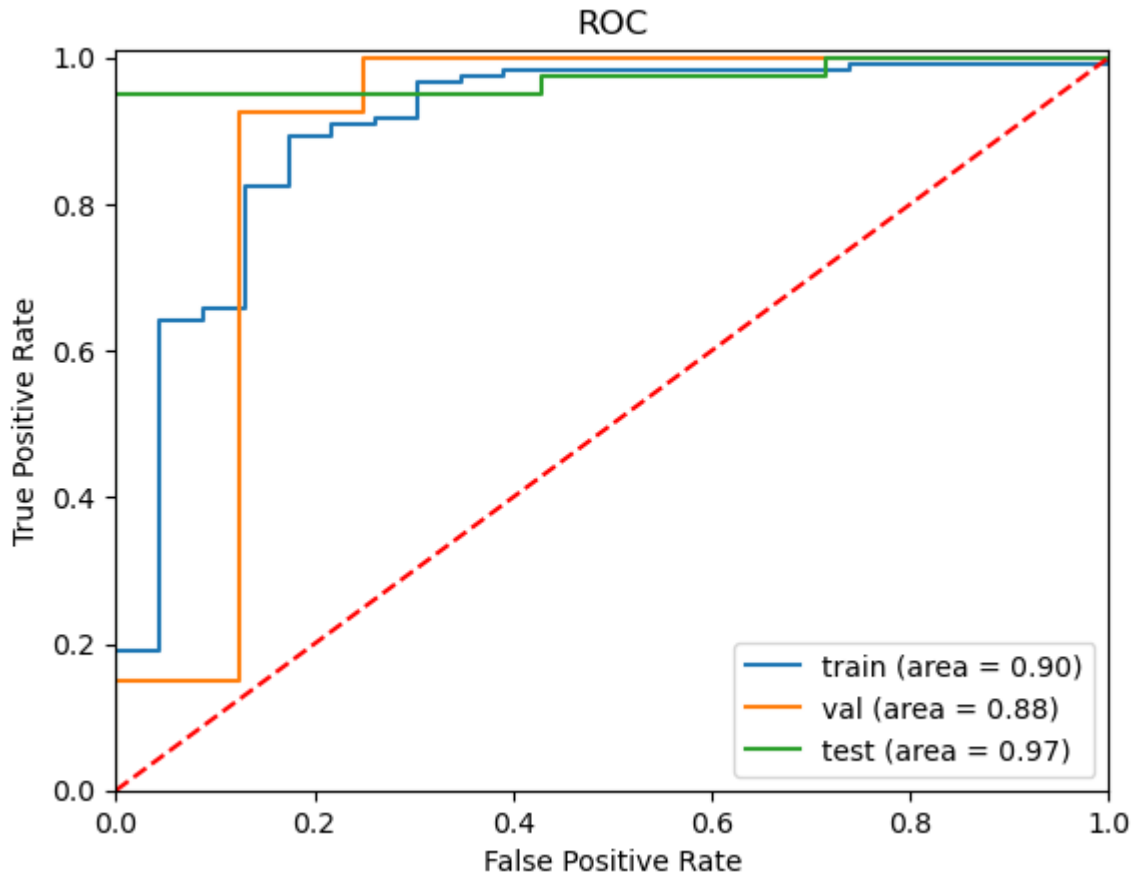
Generally, minimizing False Negatives is critical in a medical setting, hence our primary focus was on the recall metric for the performance measure. We also set the feature selection process to narrow down to only 5 key frequencies, approximately 5% of the available 109 frequencies.

The sampling techniques SMOTE and SMOTETomek yielded identical samples, leading to the same models. Overall, models trained on the sampling sets outperformed models trained on the original imbalanced sample, indicating their efficacy in handling class imbalance.

Regarding feature selection, models with features chosen by the embedded Lasso (L1) regularization demonstrated lower recall values compared to those chosen by the Ridge (L2) penalty. However, the ridge models exhibited more overfitting issues, resulting in a higher recall decrease.

Our ultimate choice is Model 5, which achieved the highest validation recall. What sets this model apart is its consistent performance on the validation set, even outperforming its training accuracy. The selected frequencies by Model 5 (890, 1029, 1414, 1587, 5993) offer valuable insights into its superior performance.

3. Discussion



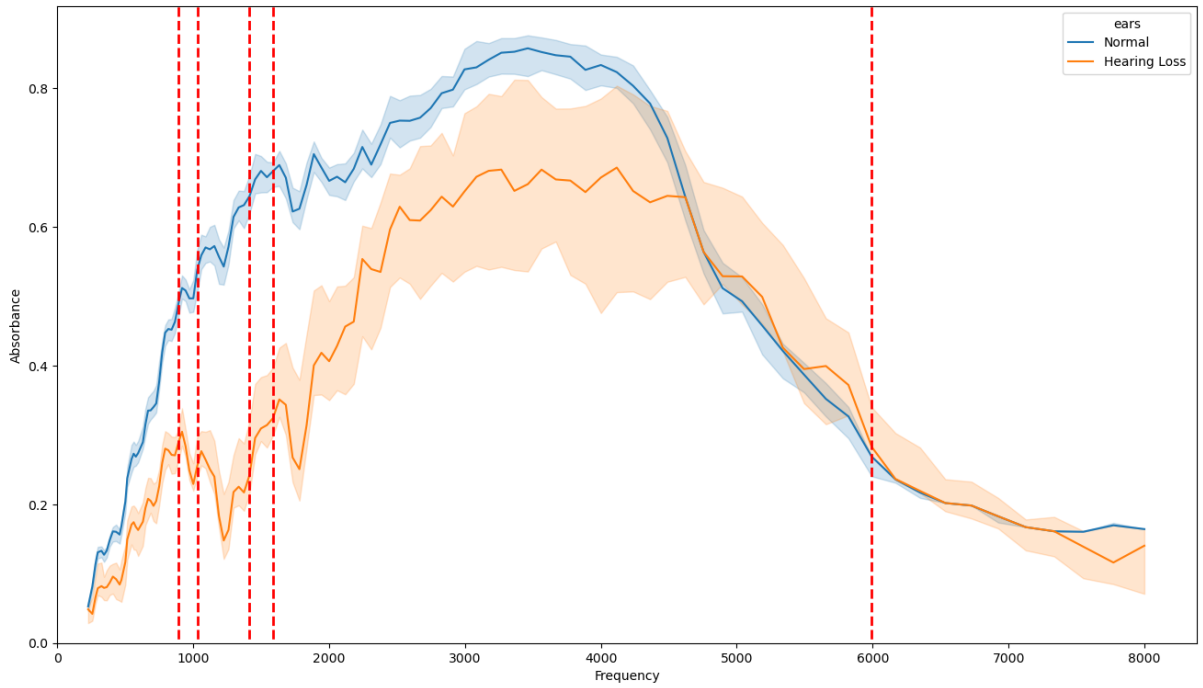
Testing Model 5 demonstrated it was able to perform consistently well on unseen data, achieving an impressive recall score of 89.02. This robust performance is further supported by its high AUC values observed across all three sets (training, validation, testing) in the ROC chart above.

To explain the model's prediction, linear SVM has the formula:
$f(X) = (w, X) + \beta = w \cdot X + \beta$, where

- $f(X)$ represents the predicted class label, with $f(X) > 0$ classified as positive and $f(X) \leq 0$ as negative.
- w is the weight vector, which is the model coefficients.
- X is the standardized absorbance levels of the frequencies
- β is the bias term, the model intercept.

Applying this formula, Model 5 has the following expression:
$f(X)$ = -0.09 + 0.2 * 890Hz + 0.22 * 1029Hz + 0.24 * 1414Hz + 0.25 * 1587Hz - 0.14 * 5993Hz.

Upon initial examination of the model coefficients, no single frequency appears to be more significantly influential compared to the others. However, it is worth noting that frequency 5993 Hz exhibits a comparatively lower coefficient factor of 0.14, whereas the other frequencies are around 0.2. Let's delve deeper into these selected frequencies:



Evidently, our final model predominantly selected frequencies in the range of 1000 to 2000 Hz, where the separation between the normal and hearing loss lines is most pronounced. We also know that the chosen frequencies within this range are more influential for the prediction compared to the 5993 Hz outside of that range. This choice likely explains why the model was able to generalize better compared to other models while maintaining a high performance.

Conclusion

Our final Linear SVM model shows promising results for accurately detecting conductive hearing loss in young children. By focusing on the absorbance levels within the frequency range of 1000 to 2000 Hz, the model effectively distinguishes between normal and hearing-impaired ears. Hence, audiologists can streamline the analysis of WBA charts in clinical settings by examining absorbance levels within that range: higher values (0.6 and above) indicate normal hearing, while lower values (0.4 and below) suggest potential hearing problems.

This frequency range (1000 to 2000 Hz) serves as an initial starting point for further studies into the role of the WBA chart in early detection of conductive hearing loss. One plausible explanation for its significance is its proximity to frequencies found in speech and other common auditory cues. However, to confirm these results, further research with a more extensive dataset is recommended for future studies.